DOCUMENT RESUME

ED 226 042 TM 830 074

AUTHOR	Mirkin, Phyllis K., Ed.; And Others
TITLE	Considerations for Designing a Continuous Evaluation System: An Integrative Review.
INSTITUTION	Minnesota Univ., Minneapolis. Inst. for Research on Learning Disabilities.
SPONS AGENCY	Department of Education, Washington, DC.
REPORT NO	IRLD-Mono-20
PUB DATE	Dec 82
CONTRACT	300-80-0622
NOTE	172p.
AVAILABLE FROM	Editor, IRLD, 350 Elliott Hall, 750 East River Road, University of Minnesota, Minneapolis, MN 55455 ($3.00).
PUB TYPE	Reports - Descriptive (141) -- Guides - Non-Classroom Use (055)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Curriculum; *Disabilities; Elementary Secondary Education; *Evaluation Methods; *Federal Legislation; *Individualized Education Programs; Program Development; Program Implementation; *Program Improvement; *Student Evaluation; Student Improvement

ABSTRACT

Federal law (PL 94-142) has charged schools with the task of constructing for handicapped students individual educational programs that specify curriculum-based goals with procedures for measuring progress toward these goals. To demonstrate substantive and procedural compliance with this law, measurement and evaluation procedures must be incorporated into the instructional program. This monograph presents a decision matrix that provides a model for developing an adequate and useful measurement and evaluation system. Available empirical work supporting the use of specific procedures for curriculum-based evaluation of student performance in reading, spelling, and written expression is discussed within the decision matrix framework. Technical, effectiveness, and logistical considerations are discussed, and data related to "what to measure" and "how to measure" decisions are provided. Alternate procedures are described for data summarization and interpretation. Technical, instructional, and logistical advantages and disadvantages of data utilization procedures are reviewed. A case study demonstrates the implementation of the recommended procedures. In this study, a teacher measured and evaluated the reading progress of a mildly-handicapped fourth grader who was reading at a second grade level. Tables and figures are appended. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Ysseldyke

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

IRLD

Director: James E. Ysseldyke

The Institute for Research on Learning Disabilities is supported by a contract (300-80-0622) with Special Education Programs, Department of Education. Institute investigators are conducting research on the assessment/decision-making/intervention process as it relates to learning disabled students.

During 1980-1983, Institute research focuses on four major areas:

- Referral
- Identification/Classification
- Intervention Planning and Progress Evaluation
- Outcome Evaluation

Additional information on the Institute's research objectives and activities may be obtained by writing to the Editor at the Institute (see Publications list for address).

The materials presented herein were prepared under government sponsorship. Contractors are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent the official position of Special Education Programs.

Monograph No. 20

CONSIDERATIONS FOR DESIGNING A CONTINUOUS EVALUATION SYSTEM: AN INTEGRATIVE REVIEW

Phyllis K. Mirkin, Lynn S. Fuchs, and Stanley L. Deno

Editors

Institute for Research on Learning Disabilities

University of Minnesota

December, 1982

Phyllis Mirkin

Those of us who worked with Phyllis for so many years to develop and improve the data-based program modification procedures described in this monograph dedicate this monograph to her memory. Her commitment to helping handicapped students through improved evaluation of instruction created a motivation for research that we all miss very much.

Abstract

This monograph presents a decision matrix that provides a model for developing an adequate and useful measurement system. Available empirical work supporting the use of specific procedures for curriculum-based evaluation of student performance in reading, spelling, and written expression is discussed within the decision matrix framework. Alternative measurement procedures for each academic area are presented, and evaluation systems applicable across areas are discussed. A case study demonstrating the implementation of the recommended procedures is presented.

Table of Contents

	Page
Chapter I: Issues in the Development of a System for Continuous Evaluation of Academic Performance Phyllis K. Mirkin	1
Context and Rationale	1
What to Measure and How to Measure	2
How to Use Data	10
Summary	13
Chapter II: Decision Framework for Developing Measurement and Evaluation Systems Stanley L. Deno, Phyllis K. Mirkin, and Lynn S. Fuchs	17
The Matrix: Broad Considerations	18
Technical Adequacy	18
Instructional Effectiveness	20
Logistical Feasibility	21
Matrix Decisions	21
What to Measure	21
How to Measure	22
How to Use Data	25
Summary	26
Chapter III: Reading Lynn S. Fuchs	29
What to Measure: The Selection of a Behavior	29
Technical Considerations	30
Construct validity	30
Concurrent validity	31
Sensitivity to student growth	36
Effectiveness Considerations	37
Logistical Considerations	37
How to Measure: The Selection of a Basic Strategy	37
Technical Considerations	38
Concurrent validity	38
Test-retest reliability	41
Effectiveness Considerations	41
Logistical Considerations	43
How to Measure: The Selection of a Score	45
Technical Considerations	45
Concurrent validity	45
Sensitivity to student growth	46
Test-retest reliability	47
Effectiveness Considerations	47
Logistical Considerations	48
How to Measure: The Selection of a Mastery Unit Within Progress Measurement	48
How to Measure: The Selection of a Difficulty Level	

Within Performance Measurement 50
Technical Considerations 50
Concurrent validity 50
Sensitivity to student growth 52
Effectiveness and Logistical Considerations 53
How to Measure: The Selection of the Size of the Measurement Domain within Performance Measurement 54
Technical Considerations 54
Sensitivity to student growth 55
Percentage of interventions resulting in student growth 56
Variability 56
Effectiveness Considerations 57
Logistical Considerations. 58
How to Measure: The Selection of a Measurement Frequency 58
Technical, Effectiveness, and Logistical Considerations 58
How to Measure: The Selection of a Sample Duration 59
Technical Considerations 59
Concurrent validity 60
Standard deviation 61
Level of performance and sensitivity to student growth. 62
Variability of time-series data 63
Effectiveness Considerations 63
Logistical Considerations 64
How to Measure: The Selection of a Mastery Criterion 64
Technical Considerations 65
Concurrent validity 65
Sensitivity to student growth 67
Congruency 68
Effectiveness Considerations 69
Logistical Considerations 69
How to Measure: The Selection of a Procedure for Generation of Test Samples 70
Technical Considerations 70
Logistical Considerations 70
How to Measure: The Selection of Test Administration and Scoring Procedures 71
Logistical Considerations 71
Summary 73

Chapter IV: Spelling 77
Gerald Tindal

What to Measure: The Selection of a Task 77
Technical Considerations 78
Validity 78
Sensitivity to student growth 81
Logistical Considerations 82

What to Measure: The Selection of a Behavior 83
Technical Considerations 83
Concurrent validity 83
Interscorer reliability 87
Sensitivity to student growth 87
Logistical Considerations 89
How to Measure: The Selection of a Scale 90
How to Measure: The Selection of a Measurement Frequency . . 91
Technical Considerations 91
Effectiveness Considerations 92
Logistical Considerations 92
How to Measure: The Selection of a Sample Duration 92
Technical Considerations 93
Validity 93
Sensitivity to student growth 93
Effectiveness and Logistical Considerations 94
How to Measure: The Selection of Sampling Procedures 95
Technical and Logistical Considerations 95
How to Measure: The Selection of Administration Procedures . 97
Technical Considerations 98
Summary 98

Chapter V: Written Expression 101
Doug Marston

What to Measure: The Selection of a Behavior 101
Technical Considerations 101
Criterion validity 102
Discriminative validity 102
Reliability 102
Sensitivity to student growth 104
How to Measure: The Selection of a Scoring Procedure 106
Technical Considerations 106
Logistical Considerations 107
How to Measure: The Selection of a Stimulus Format 107
Technical Considerations 108
Validity 108
Reliability 108
Logistical Considerations 109
How to Measure: The Selection of a Measurement Duration . . 109
Effectiveness Considerations 110
Logistical Considerations 110
How to Measure: The Selection of a Measurement Frequency . . 111
Summary 111

Chapter VI: Data Utilization 115
Lynn S. Fuchs

Graphing 116
Progress Measurement: Selecting a Graphing Convention . 116

Performance Measurement: Selecting A Graphing Convention 117
Data Summarization and Interpretation 118
Goal-Oriented Analysis 118
Program-Oriented Analysis 119
Data summarization methods 120
Data interpretation methods 120
Technical Considerations 121
Effectiveness Considerations 122
Logistical Considerations 122
Summary 124

Chapter VII: A Case Study 127
Lynn S. Fuchs

What to Measure: The Selection of a Behavior 128
How to Measure: The Selection of a Basic Strategy 128
How to Measure: The Selection of a Score, a Difficulty . . . 128
Level, and a Measurement Domain 128
How to Measure: The Selection of a Measurement Frequency and a Sample Duration 129
How to Measure: The Selection of a Criterion of Mastery or Goal 129
How to Measure: The Selection of a Procedure for Generating Test Samples 130
How to Measure: The Selection of Administration and Scoring Procedures 130
Goal and Objective Form 131
Measurement System Form 131
How to Use Data 132

References 135

Tables 145

Figures 149

Chapter I Outline

Context and Rationale

What to Measure and How to Measure

How to Use Data

Summary

Chapter I

Issues in the Development of a System for Continuous Evaluation of Academic Performance

Phyllis K. Mirkin

Context and Rationale

Federal law (PL 94-142) has charged schools with the task of constructing for handicapped students individual educational programs (IEPs) that specify curriculum-based goals with procedures for measuring progress toward those goals. To demonstrate substantive as well as procedural compliance with this law, measurement and evaluation procedures must be incorporated into the instructional program (Deno & Mirkin, 1980; Jenkins, Deno, & Mirkin, 1979). By doing so, one creates the data base with which student programs can be improved and progress toward goals can be monitored continuously.

Scriven (1967) described continuous evaluation of programs and goals as formative, wherein information is generated about the adequacy of programs while they are still in progress. In contrast to summative evaluation, which evaluates programs once they have been terminated, formative evaluation permits on-going changes to improve continuously students' programs. In formative evaluation, the concern is with program improvement rather than with final judgments about a program's efficacy.

Therefore, a critical component of formative evaluation in the IEP process is the use of information that has been gathered during program implementation to formulate decisions concerning program changes. The information-gathering process must generate data that

will verify the extent to which program changes lead to program goals (Churchman, Petrosko, & Spooner-Smith, 1975).

To incorporate formative evaluation successfully into an instructional program, three requisite decisions are (a) what to measure, (b) how to measure, and (c) how to use data to determine whether programs are efficacious.

What to Measure and How to Measure

The issue of what behavior to measure is critical in designing a formative evaluation system. Howell, Kaplan, and O'Connell (1979) stated that for a formative evaluation system to be effective it must be instructionally useful, incorporating instructional variables. The more directly the procedures measure what is taught, the more likely the measure will have instructional utility. Consequently, if reading is being taught, measurement should require children to read; if spelling is being taught, measurement should require spelling, etc. Unfortunately, many educational tests are less direct.

Test development has been dedicated primarily to the design of instruments that are associated with, rather than direct indices of, learning (Stake, 1971). The purpose of such indirect educational tests has been to demonstrate variance among human abilities, rather than to assess what knowledge a student has mastered (Gagne, 1965). In Conditions of Learning, Gagne (1965) stated:

> Despite the existence of a rather elaborate technology, it cannot be said with confidence that the assessment procedures customarily used in developing typical "standardized tests" are entirely adequate to meet current assessment needs. One important problem that does not appear to have been included...is a method for assessing human performance in terms of the objectives of instruction. (p. 258)

Gagne (1965) proposed that tests ought to "assess the immediate outcomes of learning" in order to evaluate whether "each learner has achieved the defined objective with which instructional planning began" (p. 258). While standardized norm-referenced achievement tests may be appropriate for prediction purposes or for gaining a normative perspective, they clearly do not meet Howell et al.'s (1979) or Gagne's (1965) criteria for a useful evaluation system. These instruments, then, do not have high instructional utility.

As an alternative, criterion-referenced tests have the instructional utility that is critical in a formative evaluation system (Haring & Gentry, 1976). Criterion-referenced measures are designed to provide information on the student's attainment of specific curriculum objectives and, as such, they fulfill Gagne's (1965) and Howell et al.'s (1979) criterion of directness.

In a typical scenario, a teacher may use a criterion-referenced instrument as a pretest to sample a student's behavior one time. In this way the pupil's current performance level is determined, and appropriate instructional goals are selected. Then, after a period of teaching, the test or its equivalent is re-administered as a posttest to determine whether the student has achieved the instructional goals. Such criterion-referenced measurement samples small domains of behavior at varying time intervals, usually determined by teacher rather than student behavior. Since the reliability and validity of these tests usually are unknown, we believe that measurement of this type provides an insufficient data base for judging instructional effectiveness. In order to permit on-going program evaluation so that

programs can be ameliorated in response to pupil behavior change, White (1974), among others, proposed that the ideal assessment schedule would be daily.

The issues of what to measure and how often to measure (daily, weekly, etc.) are addressed by Van Etten and Van Etten (1976) in their two-dimensional matrix of educational measurement. These two dimensions are frequency and directness. On the frequency dimension, measures are designated as continuous or non-continuous. Continuous measures are those in which a session-by-session record of change in pupil behavior is provided. Non-continuous measures are those administered at periodic intervals (e.g., every six weeks, every semester, at the completion of a unit of study). On the directness dimension, measures are characterized as either direct or indirect, based on the correspondence between what behavior is taught and what behavior is sampled by the test. Direct measures are those which test precisely the same skills as have been taught; they often use the same response mode as that employed initially in teaching the skills. Indirect measures are those in which test items usually are sampled from a larger domain and are not necessarily the items that have been taught.

For the purpose of this discussion, the model of Van Etten and Van Etten (1976) has been adapted (see Table 1). The frequency dimension is divided into frequent and infrequent rows rather than the continuous/non-continuous dichotomy, which appeared inappropriate because daily measurement is not synonymous with continuous measurement. Such an adaptation results in four measurement cells:

Type I: indirect and infrequent; Type II: indirect and frequent; Type III: direct and infrequent; Type IV: direct and frequent (see Table 1).

Insert Table 1 about here

Van Etten and Van Etten (1976) proposed that different categories of measurement may be appropriate for different purposes. The closer the measurement system is to Type IV, however, "the more direct and the more continuous the data system becomes and the more precise the teaching-learning process becomes" (p. 480). When an infrequent measurement strategy is used, even if the measures are direct (i.e., Type III), data-based decisions must await specified testing intervals. When direct and frequent measurement occurs (i.e., Type IV), data-based decisions can be made routinely during the student's program.

Proponents of the use of direct and frequent measurement (Type IV) for program development assume that no matter how careful the initial assessment might have been, educational program planners currently are unable to predict those interventions that consistently will be effective. At best, program planning decisions are hypotheses that must be tested empirically to determine their effectiveness for an individual student. Morrissey and Semmel (1976) noted that "the teacher's ability to make decisions, probably more than any other variable, affects how and what a child will learn" (p. 114). The assumption in Type IV assessment is that the more direct and frequent

the teacher's information is, the higher the probability that the teacher will make appropriate decisions, which in turn will lead to increased student achievement.

Systems termed variously as Precision Teaching (Lindsley, 1964, 1971), Exceptional Teaching (White & Haring, 1976, 1980), and Data-Based Program Modification (Deno & Mirkin, 1977) utilize direct and repeated assessment to develop, implement, and evaluate instructional programs. The methodology is an adaptation of the principles derived from operant psychology and the experimental analysis of behavior (Skinner, 1938). Rate of student performance most frequently is measured to assess the effectiveness of programs in achieving the objectives. Alper and White (1971) noted that:

> Precision teaching and operant psychology are both self-examining, self-evaluating, and self-correcting analytic paradigms which have been demonstrated through countless basic applied research projects to be both effective and easily implemented methods for the development and improvement of almost any behavior. (p. 445)

Proponents of this view suggest that direct and continuous evaluation of performance makes it possible to monitor performance and make changes in the program with the greatest benefit to the student. The available literature, however, reveals only a limited body of research in which these procedures were evaluated and contrasted with more traditional measurement practices.

One of the earliest systematic applications of repeated, direct measurement to groups of children with learning and behavior problems in a natural setting was conducted by Haring and Lovitt (1969). The study investigated the efficacy of these procedures in the remediation and prevention of academic failure. In four school districts, 55

teachers implemented over 1,250 individual plans that utilized direct and daily measurement. Unfortunately, the effectiveness of this project is difficult to assess because no data were reported on the ratio of successful to unsuccessful projects, or on overall comparisons with other treatments.

A subsequent study by Haring, Maddux, and Krug (1972) provided experimental data. Twenty-four children received reading and math instruction under "systematic" precision teaching procedures that included direct and repeated measurement of academic performance. A contrast group received instruction under procedures typically available in special classes. The performance of these two groups was contrasted. Pre-posttest gain scores on the WRAT revealed that, over an eight-month period, the experimental group made average gains of 13.5 months in reading and 16.0 months in math, while the control group made gains of 4.5 months and 5.0 months, respectively, in reading and math.

In a one-year follow-up study, Haring and Krug (1975) compared the achievement of the experimental subjects who had returned to regular classrooms with regular class students who had been matched with the experimental group on reading scores. Students were tested on two standardized achievement tests and nine academic and nine social adjustment skills. Teachers ranked high numbers of both groups in the middle third of the class in math and the majority of students in the upper two-thirds of the class socially. Additionally, 76% of the experimental and 61% of the matched students were ranked as capable of remaining in the regular classroom while the remainder were

judged as needing some outside assistance but not special class placement. In an eight-month period, the experimental group, on the average, gained 13 months in reading and 9 months in math on the WRAT, while the matched subjects gained 7 months and 6 months, respectively. While there are methodological inadequacies in these studies, they represent an important attempt to validate the effect on school performance of a highly structured social and academic program that included direct and daily measurement.

Frumess (1973) investigated daily measurement and different degrees of self management in a study with 45 boys randomly selected from self-contained classrooms for minimally brain-injured (MBI) students.[1] Within three classrooms, students were assigned to one of five groups: (a) Self-Chart, Self-Set Aims (SCSSA)--Students graded, tallied, and recorded their own frequencies on math fact performance and set their own weekly aims on a graph. Aims were reviewed daily but teachers did not influence students regarding what aim to set; (b) Self-Chart, Teacher Set Aims (SCTSA)--This group followed exactly the same procedure as SCSSA except that the teacher set the weekly aim. Aims were set in decreasing order, from ten facts per minute to two facts per minute more than the student's score on the previous Monday for each of the eight weeks of the study (i.e., if a student scored 30/min correct on Monday, the aim for the first week was set at 40/min. If on the following Monday the score was 32 correct/min, the aim for Friday would be 41 correct/min, etc.); (c) Teacher Chart, Teacher Set Aims (TCTSA)--This group also graded and tallied the number of math facts correct/min and incorrect/min, but this was their

only feedback. Teachers charted and set weekly aims. Charted information could be used to plan instruction, but teachers did not give students any information on their rate of improvement or make comparisons between students' charts; (d) No Charting or Setting Aims (NCSA)--These children graded and tallied number of math facts correct and incorrect daily but they neither charted nor set aims; (e) Control Group--This group was provided traditional daily arithmetic instruction by their teacher.

Although the SCSSA and the SCTSA groups showed greater gains from pre to posttesting than all other groups, there were no reliable differences in performance between these two groups. The NCSA group made significantly greater gains than the control group and the TCTSA group. Of particular interest is the significantly greater improvement of the NCSA group over the TCTSA group. Frumess noted that it is unclear how the teachers used the data for purposes of instructional decision making, and that her findings possibly are the result of inadequate data utilization or lack of data utilization, a problem that has caused considerable consternation among advocates of direct and daily measurement.

In a review of individual data projects implemented at various public schools in the Washington area, White (1971) found that many teachers allowed programs to remain in effect long after their usefulness had been exhausted. Although teachers collected data, they did not use those data as a signal to change programs when goals or objectives were not being achieved at the desired rate. This raises the third major issue that must be resolved when developing a

formative evaluation system; that is, how to use the data once they have been collected.

How to Use Data

Initial attempts to provide "rule-of-thumb guidelines" for data utilization (Haring & Lovitt, 1969) required (a) analysis of rate accelerations and decelerations to determine when a program change should be made; and (b) simultaneous analysis of the slope of the data, medians, step changes at point of intervention, and variability. Data-utilization rules that provide standard methods for daily program analysis also have been developed (Haring, White, & Liberty, 1979; Liberty, 1972, 1975; White, 1971). These rules attempt to remove the "guesswork" from data analysis by providing guidelines with respect to the length of time interventions should be held constant.

An investigation by Bohannon (1975) compared the achievement of children whose programs were directed by daily measurement and decision rules (experimental group) with that of children whose treatment was directed by teacher judgments only (control group). Visual analysis of average gain scores revealed that students in the experimental treatment gained three to four times more than students in the traditional treatment (control) group in reading phonic words. Twenty-two of the twenty-three experimental students scored within the interquartile range of normally achieving mainstream peers for reading phonic words. Only 3 of 23 traditionally treated children met these criteria. Additionally, 8 of 23 children required only daily assessment to maintain sufficient progress to achieve the aim. Two children required one program modification, eight required two

modifications, and two required a third intervention before achieving their aims. Only 3 of the 23 children did not achieve their aims in the specified time period but these children did achieve their highest rates in the final phase. Reliable differences favoring the experimental group also were obtained on seven of the eight items of an affective development questionnaire completed by regular classroom teachers. Interestingly, the experimental group achieved these results in approximately one-third the daily instructional time allotted to the traditional treatment group.

Additional data on the effects of repeated assessment and data decision rules on academic achievement appear in the work of Mirkin and Deno (1979). These researchers contrasted daily oral reading practice, daily oral reading practice with measurement, and daily practice with measurement and specific data-utilization decision rules. They found reliable differences that favored the latter group. Unfortunately, the data utilization treatment included rules for changing goals as well as delivery of consequences. Consequently, the results were confounded.

Deno, Chiang, Tindal, and Blackburn (1979) subsequently compared the effects of data-utilization rules that allowed students to earn points when they achieved or exceeded their daily aim with data-utilization rules that included feedback on performance and praise when the daily aim was reached. Reliable effects favoring the data-decision rule with points were found. In an additional analysis, the effect of using graphs with an aimline was contrasted with graphing without an aimline. The results revealed a higher proportion of

students achieving criterion when daily performance was graphed without aimlines.

Mirkin, Deno, Tindal, and Kuehnle (1980) examined teacher effectiveness in spelling programs when (a) performance was measured weekly and teachers used their judgment to make program changes, (b) performance was measured daily and teachers used their judgment to make changes, and (c) program changes were based on daily measurement and specific decision rules. Reliable differences favored daily measurement over weekly measurement, but no reliable difference between daily measurement conditions (teacher judgment vs decision rules) was revealed.

Martin (1980) examined the effects of variations in data utilization on the reading performance of mildly handicapped students. The data-utilization procedures included a decision rule determining when to change an instructional program and another determining both when and what to change. Forty-five mildly handicapped students were assigned randomly to one of three conditions: Group 1 ("When and What" Decision Rules) with changes made in accord with "Experimental Data-Decision Rules with Minimum Celeration" (Haring et al., 1979); Group 2 ("When" Decision Rules) with changes selected by teachers; and Group 3 (No Rules) with daily correct and error rate recorded but not charted. Results revealed that all students improved during the experimental period. Statistically significant differences favored Group 1 over Group 3 in the proportion of students achieving established aims. At the instructional level, the posttest performance of Group 1 students was significantly less than Group 3

students on reading-in-context, but greater than Group 3 students on comprehension tasks at reading levels beyond their instructional placements.

Summary

Results of experiments on direct and frequent measurement and evaluation generally support its effectiveness in improving student achievement. Nevertheless, available research on such measurement and evaluation strategies is replete with differences in the types of behaviors measured, the data recorded, the data-decision rules employed, and the types of changes introduced. These differences across studies along with the confounding treatments within experiments render it difficult to draw conclusions concerning which aspects of a measurement and evaluation system are critical in affecting student achievement. Further, there is scant information on the edumetric adequacy or the logistical feasibility of such procedures.

Given this situation, it is difficult for a practitioner to determine what behavior to measure, what measurement methodology might be appropriate, and how to use data once they are collected. Further, how to ensure that the resulting measurement and evaluation system will positively affect student behavior, be technically adequate, and be logistically feasible to implement in the classroom is unknown.

Consequently, this monograph has two purposes. The first is to describe a model within which one might design a direct and frequent measurement and evaluation system. The second purpose is to discuss, within the framework of the model, the available empirical work that

supports the use of specific procedures for repeated curriculum-based evaluation of student performance in reading, spelling, and written expression.

The monograph is organized as follows. Chapter II describes a model for developing an adequate and useful measurement system. In Chapters III, IV, and V, this model is employed to discuss alternative measurement procedures in three academic domains, while in Chapter VI different evaluation systems are discussed. Finally, in Chapter VII, a case study demonstrating the implementation of the recommended procedures is presented.

Chapter II Outline

The Matrix: Broad Considerations

- Technical Adequacy
- Instructional Effectiveness
- Logistical Feasibility

Matrix Decisions

- What to Measure
- How to Measure
- How to Use Data

Summary

Chapter II

Decision Framework for Developing Measurement and Evaluation Systems

Stanley L. Deno, Phyllis K. Mirkin, and Lynn S. Fuchs

As suggested in the preceding chapter, the learning principles of educational psychology and applied behavior analysis provide a theoretical rationale for incorporating measurement and evaluation into instruction. The merger between performance monitoring and instruction is not only conceptually sound, but also supported by research investigating the impact of systematic evaluation on student academic achievement (see Chapter I), and mandated by Federal law (PL 94-142).

Given compelling arguments for integrating instruction and measurement, what remains is to create measurement and evaluation systems that (a) can be incorporated efficiently into instructional procedures, (b) satisfy technical edumetric requirements, and (c) result in improved student achievement. Toward that end, the purpose of Chapter II is to present a decision framework for developing satisfactory measurement and evaluation systems.

Table 2 displays the decision-making matrix that was adopted for this discussion. The matrix rows contain three broad requisite decisions in the development of measurement and evaluation systems: what to measure, how to measure, and how to use data. The matrix columns list three broad considerations in specifying these procedures: technical adequacy, instructional effectiveness, and logistical feasibility. Within the matrix, a question corresponding

to the intersection of each decision and each consideration is provided (see Table 3). This decision-making matrix was designed to guide the developer of a measurement and evaluation system; it is examined in the remainder of this chapter.

Insert Tables 2 and 3 about here

The Matrix: Broad Considerations

Technical Adequacy

Some technical adequacy issues that are important in developing a formative evaluation system are similar to those relevant to traditional, norm-referenced assessment (validity, reliability). Others are unique to the concerns of time-series measurement.

The relevant technical adequacy issues of norm-referenced tests include the validity of the direct measure of student achievement, and the measure's reliability. Validity refers to the extent to which a specific measurement procedure produces data directly related to the purposes of measurement. Used in this way, the validity of a test will depend upon the degree to which it improves decision making. Three types of validity have been identified that relate to the major purposes of measurement. They are criterion validity, content validity, and construct validity. In order for a measure to demonstrate strong criterion validity, a student's score should correlate highly with other data deemed important. For example, for a reading measure to have criterion validity it should correlate with technically adequate standardized tests of reading, teacher placement

in the curriculum, or placement in special reading programs. If we explore "reading aloud from text," low scores on reading aloud from text should be associated with poor performance on standardized reading tests, while high scores should be characteristic of students who perform well on the standardized tests. Content validity is dependent on the adequacy with which a specified domain of content is sampled. For example, reading aloud from text would demonstrate content validity if experts agreed that the task represented what a student should learn to do in reading. Finally, construct validity represents the extent to which measures correlate in expected ways with other measures or are affected similarly by experimental treatments. Once a domain of behaviors is specified along with the ways in which those behaviors relate to one another, experimental evidence is sought to confirm or disconfirm that those behaviors actually do relate to one another as hypothesized.

A second relevant technical adequacy issue is reliability, including test-retest reliability, alternate-form reliability, and interscorer agreement. Test-retest reliability is illustrated in the following example. If John reads words from the same third grade level word list on Monday and then on Wednesday, we would expect his performance to be very similar. The extent to which the scores are alike is an indication of the test's test-retest reliability. Alternate-form reliability is highly relevant for repeated measurement, where one employs a different form of a test at each measurement session. The extent to which a student's scores are similar on alternate forms administered on the same day is an

indication of the test's alternate-form reliability. Interscorer agreement refers to the extent to which a student's performance is scored alike by two or more independent examiners. All forms of reliability indicate the extent to which scores on a test are free from error and represent a student's "true score."

In addition to these traditional technical concerns, an ongoing measurement system must consider the following technical issues that relate more directly to time-series measurement. One of these concerns is slope, which is indicative of a measure's sensitivity to student growth or the extent to which student achievement is manifest in performance on a measure. Measures for which there are greater slopes or ranges of behavior change over time better allow small amounts of growth to be registered. For example, if a first grader were reading a book where the average change over a year for a first grader was 15 words, the measurement would be less sensitive to student growth than if that first grader were reading another book where the average change over a year for first graders was 65 words.

Other technical concerns include: (a) the amount of intra-individual variability, that is, within an individual's graph, how much variability exists between adjacent data points; (b) the degree to which the measurement behavior is linked directly to the IEP, the child's curriculum, and the terminal behavior (see Van Etten & Van Etten, 1976; "direct" measures); and (c) the reliability or consistency of teachers' interpretations of data.

Instructional Effectiveness

The purpose of an on-going measurement and evaluation system

ultimately is to improve the effectiveness of an instructional program. In selecting among elements, it is preferable, of course, to choose those that improve teacher decision making and maximize student growth. Therefore, in addition to technical concerns, selection among measurement and evaluation components must reflect the effects those elements have on teacher decisions and student achievement. Instructional effectiveness, consequently, is included in the decision-making matrix.

Logistical Feasibility

Frequent measurement appears to be time consuming for teachers to implement. In research conducted in a rural, special education cooperative (Fuchs, Wesson, Tindal, Mirkin, & Deno, 1981), elementary resource teachers initially spent an average of 2 1/4 minutes preparing for, administering, scoring, and graphing one measure for one student. Multiplied across a full caseload of students, this figure represents a large portion of teacher time. Therefore, in designing a feasible measurement system, one must make logistical changes to reduce teacher and student time in measurement.

Matrix Decisions

What to Measure

In developing a technically adequate set of measurement procedures, one must first determine the measurement behavior, that is, the specific skill to be quantified. For example, in reading, one might consider reading isolated words aloud or silently, reading text aloud or silently, answering questions based on text reading, decoding nonsense words, or completing cloze passages. In spelling, one could

explore writing from dictation, editing word lists, recognizing correct alternatives, or writing self-generated paragraphs. In other words, one must select the specific behavior or behaviors that will be measured in a global area such as reading or spelling.

How to Measure

The decisions constituting "How to Measure" vary by academic domain. Figures 1 and 2 illustrate the decision flows for reading and spelling, and for written expression, respectively. The following discussion briefly describes each decision referenced in Figures 1 and 2.

Insert Figure 1 and 2 about here

The selection of a basic measurement strategy for monitoring student progress is an essential decision in the design of a measurement system. Two alternate strategies are performance and progess measurement (Deno & Mirkin, 1977). Performance measurement provides information on how a student's behavior changes on a task of constant difficulty. In performance measurement, increases in fluent performance on equivalent forms of the task should represent growth or ...evement. For example, a teacher might decide to measure a student's performance on reading aloud from a fourth grade reader. ...h day, the teacher would select randomly a passage to measure student performance and ask the student to read aloud for one minute. Within this strategy, the student's graph might display the number of words correct and the number of errors per day in one minute of

reading from the fourth grade text.

A second strategy is progress measurement. As illustrated in the work of Deno and Mirkin (1977), progress measurement involves monitoring student mastery through a curriculum over a period of time. In progress measurement, a sequence of objectives is specified and a criterion of mastery established for each objective. Mastery of objectives then is assessed frequently to monitor student progress. For instance, a teacher might establish a series of phonics skills as the sequence of objectives to be mastered. Then the teacher might determine that a performance standard of reading 50 words per minute correctly with no more than two errors be the criterion of mastery that is to be met before a student can progress to the next objective. The student's graph, therefore, would display objectives mastered per time unit, and improved progress would be indicated by an increased rate of mastery through the objectives.

Consequently, performance and progress measurement are different in two essential ways: (a) In performance measurement, the measurement task is sampled constantly from the same pool of material; in progress measurement, the measurement task changes each time the student masters a segment of the curriculum; and (b) in performance measurement, the goal is to describe changes in performance on one specific level of material; in progress measurement, the object is to describe the rate of progress through a series of tasks (see Table 4).

Insert Table 4 about here

Within each measurement strategy, one must select a score (correct rate, percentage correct, or incorrect rate) to employ in analyzing the measurement sample. Depending on the level of student behavior, one might want to score small units of behaviors such as letter sequences spelled correctly, or large units of behavior, such as words spelled correctly. Depending on the teacher's time available to score, one might want to analyze types of phonetic errors made or just number of errors.

Within performance measurement two additional decisions remain. First, one must determine <u>at what difficulty level measurement will occur</u> (at instructional level, at age-grade appropriate material, etc.). This difficulty level remains constant as the student's proficiency changes. One must also select the <u>size of the measurement domain</u>; that is, given a difficulty level of material, one must determine the size of the material pool from which frequent measurement tasks will be sampled (from several grade levels of material, within one grade level of material, within one unit of material, etc.).

Within progress measurement, one additional decision remains. A <u>unit of mastery</u> (pages, stories, units, books in reading; words, lists, units, books in spelling) must be determined. Often, this is problematic because curricula are not designed so that mastery units are equivalent, a requirement for a technically adequate measurement

system.

Once the basic measurement is established, a measurement frequency must be determined; that is, will measurement occur daily, twice per week, weekly, monthly, etc. Additionally, one must establish student mastery criteria. Within progress measurement one must determine mastery criteria for each step or objective within the hierarchy of skills. Within performance measurement, one must determine criteria that specify, in terms of the long-range goal or outcome behavior, when acceptable performance has been met.

Three remaining decisions involve (a) procedures for generating test samples, (b) procedures for test administration, and (c) determining how long each measurement sample will last--30 seconds, one minute, three minutes, etc. These parameters specify the mechanics of measurement. These mechanics must be outlined and held constant if the measurement data are to be interpreted meaningfully. Generation of test samples refers to how the numerous equivalent measurement samples will be created. Administration of test samples relates to the standard procedures (directions, setting, schedule) employed in administering the measures. Duration of test samples addresses how long a test lasts during each administration.

How to Use Data

The decisions grouped under the label "How to Use Data" are as follows:

(a) Data summarization methods--statistics are used to collapse data into a form in which they are interpreted easily and meaningfully.

(b) Graphing conventions--type of graph paper and procedures used to display data.

(c) Data-utilization procedures--the rules under which measurement data are interpreted, student performance is evaluated, and program changes are determined.

Summary

This chapter described the decision-making matrix for developing measurement and evaluation systems. The next three chapters (III, IV, and V) will discuss available research in the domains of reading, spelling, and written expression, respectively. Each chapter gives consideration to technical, effectiveness, and logistical considerations and provides data related to the "What to Measure" and "How to Measure" decisions. Because "How to Use Data" decisions cut across academic domains, they are discussed separately (Chapter VI). In each chapter, relevant questions posed in the matrix (see Table 3) are discussed and research-based recommendations are made.

Chapter III Outline

What to Measure: The Selection of a Behavior

How to Measure: The Selection of a Basic Strategy

How to Measure: The Selection of a Score

How to Measure: The Selection of a Mastery Unit Within Progress Measurement

How to Measure: The Selection of a Difficulty Level Within Performance Measurement

How to Measure: The Selection of the Size of the Measurement Domain Within Performance Measurement

How to Measure: The Selection of a Measurement Frequency

How to Measure: The Selection of a Sample Duration

How to Measure: The Selection of a Mastery Criterion

How to Measure: The Selection of a Procedure for Generation of Test Samples

How to Measure: The Selection of Test Administration and Scoring Procedures

Summary

Chapter III

Reading

Lynn S. Fuchs

In the preceding chapter, a decision-making matrix for selecting measurement and evaluation procedures was developed. The rows of this matrix list the broad parameters of a system; the columns list the major decision-making criteria for selecting among alternatives within a parameter. For each matrix cell, a question was posed that needs to be answered to specify a measurement system. The purpose of Chapter III is to answer each question for the area of reading. In this chapter each system parameter is discussed separately. For each parameter, (a) aiternative procedures are reviewed briefly, (b) research results as they apply to each decision-making criterion are presented, and (c) conclusions are drawn concerning those procedures that appear most useful in light of inferences and available research.

What to Measure: The Selection of a Behavior

Because direct repeated measurement is, by definition, implemented frequently, if not daily, it requires a significant time commitment from the teacher. Consequently, the behavior to be measured must meet certain logistical criteria. Important practical considerations mentioned by Deno, Mirkin, Chiang, and Lowry (1980) are ease in administration, availability of many parallel forms, cost and time efficiency, and unobtrusiveness with respect to routine instruction. The search is for behaviors that simultaneously demonstrate feasibility, instructional utility, and technical adequacy (including construct validity, concurrent validity, and sensitivity to

growth). Studies relevant to the identification of reading behaviors with these characteristics are reviewed here.

Technical Considerations

Construct validity. Studies investigating the construct validity of various reading behaviors have sought to identify the specific skills of reading comprehension and to demonstrate that these various skills exist as differential characteristics of students. Traxler (1941) studied the performance of tenth-grade students on the Wagenen-Dvorak Diagnostic Examination of Silent Reading Abilities to ascertain whether the different sections of the test yielded information sufficiently independent to warrant their separate use. He found the parts to be highly intercorrelated, and concluded that the separate parts did not contribute anything greatly different from the test's total reading score. In fact, when the intercorrelations were corrected for attenuation, most approached unity.

On the basis of a comprehensive literature survey, Davis (1944) listed nine possible categories of basic skills of reading comprehension. He then developed test questions to measure these skills, administered the tests to subjects, and computed intercorrelations among the nine tests. Following a factor analysis of the results, Davis concluded that six factors were significantly reliable for practical use. However, Thurstone (1946) reanalyzed Davis' data employing a different factor analysis technique; he concluded that a single factor was sufficient to account for the obtained correlations. Davis (1946) continued to maintain that his six factors represented significant dimensions of reading

comprehension, but he admitted that several of them accounted for very little variance in reading scores. Conant (1942) developed tests to measure general reading comprehension as well as specific reading skills, and administered them together with the Nelson-Denny Reading Test and American Council Psychological Examination. Intercorrelations among all the measures were above .50, leading Conant to conclude that the results were explained largely by a single factor, tentatively defined as general comprehension. Other factor analytic studies have corroborated this pattern of results (Artley, 1944; Hall & Robinson, 1945; Stoker & Kropp, 1960). These studies suggest that reading comprehension, as presently understood, is a general ability that is not segmented reliably into specific differential subskills.

Concurrent validity. Concurrent validity studies examine the usefulness of a measure in predicting performance on other variables. Typically, one is interested in assessing the suitability of substituting a test for a longer, more cumbersome, or more expensive criterion that has demonstrated technical adequacy (Messick, 1980). Criterion-relatedness is determined by correlational analysis, where the strength of a correlation between two measures specifies the degree of predictive efficiency between the tests (Nunnally, 1978). Therefore, in studies of concurrent validity between simple measures and standardirzed achievement tests, if correlations are high, then predictive efficiency is demonstrated between the tests. On that basis one can assume that (a) simple tests demonstrate the validity of and represent the same constructs as the longer, more global

achievement tests; (b) the simple tests provide information on students' standings relative to the normative population on which the criterion tests were standardized; and (c) as a student manifests improvement on the simple measure, his/her standing relative to that of the norm group also has improved. Therefore, concurrent validity bears directly on the technical adequacy of simple, direct reading measures. A brief review of relevant studies follows.

One group of concurrent validity studies investigated the relationship between reading comprehension and reading rate. The results of these studies were contradictory. Judd (1961) compared rapid, medium, and slow reading with good, medium, and poor reading quality. He found a negative correlation between rate and comprehension. Eurick (1930), on the other hand, found positive but weak (average $\underline{r}$ = +.31) relations between several rate measures and scores on comprehension tests. In a review of pertinent studies, Gates (1921) reported great inconsistency, with correlations between rate and comprehension ranging from -.14 to +.92.

In another summary of investigations relating to this issue, Gray (1925) agreed that the correlation between speed and comprehension was variable. However, he attempted to analyze that pattern of variability and found that the strength of the relationship was dependent on the age of the subjects. In more recent research corroborating Gray's conclusion, Sassenrath (1972) revealed that while reading rate and comprehension were separate factors at the college level, they fused into one factor at the elementary and high school levels.

In addition to student age, the nature of the material read has emerged as an important factor affecting the relationship between speed and comprehension. When easy narrative prose is employed to measure reading rate and difficult material is used for measuring the percentage of correctly answered reading comprehension questions, the correlations between accuracy and speed are low (Tinker, 1929). In contradistinction, when rate and comprehension are measured on identical reading text, the correlations consistently are high (Tinker, 1939). This finding is not surprising because the strength of the relationship even between two comprehension measures (Gates, 1921; Pressey & Pressey, 1921) or between two oral reading rate measures (Gates, 1921; Patterson & Tinker, 1930; Tinker, 1929) is affected similarly by varying the nature of the material read. Gates (1921) buttressed the view that, given comparable material, rate and accuracy are related significantly and positively; he found that composite scores of reading rate correlated strongly with composite scores of comprehension (r = +.84). Therefore, it appears that a significant relationship exists between reading comprehension and reading rate when the students are of elementary age and when the nature of the material read is held constant.

This first set of concurrent validity studies revealed a significant relationship between measures of reading comprehension and speed, and suggested that simple measures of reading rate and/or comprehension can be used to monitor the development of general reading skills. A second set of concurrent validity studies focused on the validity of simple, direct behaviors with respect to

standardized achievement tests.

Studies have investigated the concurrent validity of phonics measures, cloze procedures, in-context measures, and isolated word tests. In one of two studies that examined the concurrent validity of phonics measures, Quilling and Otto (1971) attempted to determine whether growth on the Wisconsin Design phonics objectives correlated with growth on a reading achievement test. Results in two schools were inconsistent, with positive correlations in one school and negative correlations in the other. Two additional factors render these inconsistent results even more difficult to comprehend: (a) Quilling and Otto employed gain scores as their dependent variable, and (b) they neglected to apply statistical tests to their data.

Askov, Otto, and Fischbach (1971) tested the same hypothesis within an improved design. These researchers found a statistically significant, positive relationship between word attack skill growth and reading achievement test performance. The report, however, failed to reveal which achievement test was employed and the strength of the significant relationship. Without this critical information, it is impossible to determine whether the statistically significant relationship reported represents an important one.

Therefore, validity studies that have attempted to document the concurrent validity of phonics measures with respect to achievement tests measuring global reading ability have been limited in number, weak in design, and inconsistent in results. The concurrent validity of simple phonics measures remains open to question.

Guszak (1969) investigated the validity of a cloze procedure with

respect to the Botel Reading Inventory (BRI) and the Metropolitan Achievement Test (MAT). Findings revealed low correlations (+.11 to +.18) between the BRI and the cloze tests with a sample of students in grades 4-6. Additionally, there were no significant relations between performance on the cloze tests and the MAT scores.

Deno, Mirkin, Chiang, and Lowry (1980) addressed the concurrent validity of reading in context measures, reading in isolation measures, and cloze procedures with respect to the Stanford Diagnostic Reading Test, Reading Comprehension subtest (Karlsen, Madden, & Gardner, 1975), the Woodcock Reading Mastery Tests, Word Identification subtest (Woodcock, 1973), and the Peabody Individual Achievement Test, Reading Comprehension subtest (Dunn & Markwardt, 1970). This series of studies found that, across correct scores, all three types of measures demonstrated concurrent validity, but that reading aloud from text (average r = +.81) and words in isolation (average r = +.83) measures consistently correlated higher with the standardized tests than did the cloze (average r = +.75).

Other studies have corroborated the criterion-relatedness of oral reading measures (Botel, 1968, Bradley, 1978; McCracken, 1962; Oliver & Arnold, 1978.) Also, in a study of a simple reading in context measure, Biemiller (1970) placed first grade students in one of eight passages at the level immediately below that in which students read with 75% accuracy. Then children were ranked according to the percentage of errors on the most difficult passages they read at criterion. A rho coefficient of +.89 between these oral reading performance ranks and scores on the MAT Reading Comprehension subtest

was reported; using Nunnally's standard (1959), this correlation provided evidence for concurrent validity.

Therefore, greater evidence exists for the concurrent validity of reading aloud measures than for either cloze or phonics measures. On this technical basis one might select reading aloud of either words in isolation or words in context as a reading behavior to measure.

Sensitivity to student growth. As presented here, slope is the average behavior change per grade level. Measures on which students manifest a steep slope and a relatively large range of behavior provide greater opportunity for students to register relatively small gains. A steep slope that results in heightened sensitivity to student growth is a desirable characteristic of repeated measurement.

Deno, Mirkin, Chiang, and Lowry (1980) compared the mean number of correct responses per grade level for reading aloud from text, for reading isolated words, and for a cloze measure. Using a split middle trend lind (White, 1971) to calculate slope, this analysis revealed flatter slopes for the words in isolation measures (average slope = 11.4) and the cloze measures (average slope = 2.2) than for reading aloud from text measures (average slope = 12.1). This pattern of results was corroborated by a similar study (Deno, Marston, Mirkin, Lowry, Sindelar, & Jenkins, 1982) when behavior was sampled across time and average scores were plotted over intervals spanning one year. Therefore, the technical evidence cited above supports the use of reading aloud from text as a behavior with which to measure reading performance.

Effectiveness Considerations

Measuring reading aloud from text presents several instructional advantages over phonics, cloze, or isolated word reading measures. First, of the four measures, reading aloud from text most closely constitutes the global reading task; therefore, its face validity is highest. Second, by listening to a student's text reading, a teacher can simultaneously assess a student's progress and analyze error and fluency patterns for instructional planning purposes. Finally, there has been some speculation (Samuels, 1981) that a teacher might assess reading comprehension by evaluating the prosodic features of a student's oral reading.

Logistical Considerations

Measuring oral reading from text appears to be easier to implement than phonics, cloze, or isolated word measures, because (a) the students read directly from books, eliminating the need for teachers to prepare word lists or to create cloze tests, and (b) reading aloud from text allows teachers to use a simple page number selection procedure (see Mirkin, Deno, Fuchs, Wesson, Tindal, Marston, & Kuehnle, 1981). We conclude, then, that reading aloud from text is a technically adequate, instructionally useful, and logistically feasible behavior to measure in a continuous evaluation system.

How to Measure: The Selection of a Basic Strategy

The selection of a basic strategy is esssential in the development of a system for monitoring student reading achievement. Specifically, one must choose between performance and progress measurement. In performance measurement, the difficulty level of the

measurement task is consistent. The measurement task is a random sample of items from a large pool of material, and the goal is to improve the level of performance on that material. In progress measurement, the measurement pool changes each time a student masters a segment of the curriculum, and the object is to increase the rate of progress through the segments of the curriculum. In performance measurement, a student's growth represents performance level per day on the same pool of material; in progress measurement, a student's growth graph displays cumulative objectives in the curriculum mastered per day. In selecting between these basic strategies, the decision-making criteria again are technical adequacy, instructional effectiveness, and logistical feasibility. The following discussion presents research data related to each of these considerations.

Technical Considerations

Concurrent validity. In a series of studies, Deno, Mirkin, Chiang, and Lowry (1980) examined the concurrent validity of simple performance reading measures. In the first study, 33 randomly selected students in grades 1 - 5 were tested on the following achievement tests and one-minute sixth-grade performance measures: reading aloud from text, reading words in isolation, a cloze test, a word-meaning test, the Stanford Diagnostic Reading Test, subtest five (Part A, Reading Comprehension) of Form B (Karlsen et al., 1975), and the Woodcock Reading Mastery Tests, Word Identification and Word Comprehension Tests of Form A (Woodcock, 1973). Results indicated that all measures correlated significantly with each other. Correlations between the word recognition measures and the achievement

tests ranged from +.73 to +.91, with most coefficients in the +.80s. Correlations between the cloze and word meaning measures (where word meaning tests asked children to define words in context) and the achievement tests ranged from +.60 to +.83. In a partial replication study, this pattern of results was corroborated with 66 subjects tested on analogous third grade performance measures and on the Phonetic Analyses and the Reading Comprehension subtests of the Stanford Achievement Test and the Reading Comprehension subtest of the Peabody Individual Achievement Test (Dunn & Markwardt, 1970).

Fuchs (1981) also explored the concurrent validity of simple reading performance measures. To 91 subjects, examiners administered a measure of reading aloud from passages, from each level in two reading series, and the Woodcock Reading Mastery Tests (Form A), Word Identification and Passage Comprehension subtests (Woodcock, 1973). Correlations of both the correct rate and the percentage correct measures with the achievement tests ranged from +.64 to +.96, with 53% of the correlations at +.90 or higher.

The high correlations between performance and achievement tests demonstrated in these three studies indicate that the performance measures of reading aloud from text are highly predictive of scores on these standardized achievement tests. On that basis one can assume that scores on these reading performance measures provide much of the same technically adequate information as do scores on the reading achievement tests employed in these studies.

In a similar fashion, Fuchs and Deno (1981) examined the concurrent validity of simple curriculum-based progress measures of

reading with respect to standardized achievement tests. On the basis of performance in reading aloud from text, 91 students in grades 1-6 were assigned seven mastery level scores according to seven different mastery criteria in two reading series. The Word Identification and Passage Comprehension subtests of the Woodcock Reading Mastery Tests (Form A) also were administered to these students. For both reading series, correlations between the mastery level scores and the achievement tests were in the +.80s and +.90s for six of the mastery cirteria and in the +.60s for the other criterion of mastery. These high correlations demonstrate the criterion validity of progress measures with respect to achievement tests. However, this study validates progress measurement of gross units, that is, books rather than stories or even pages. It is unclear whether mastery stories within a book would correlate with improved performance on standardized achievement tests.

Nevertheless, available evidence supports the concurrent validity of both progress and performance measures with respect to achievement tests. Both progress and performance measures based on reading aloud from text appear to demonstrate the validity of, and represent the same reading constructs as, longer and more global achievement tests. Both appear to provide information on students' standings relative to the normative groups on which the achievement tests were standardized; further, improvement on either type of measure represents improvement in a student's standing relative to the achievement tests' normative populations. Consequently, concurrent validity studies support the utility of performance and progress measures as a substitute for

longer and more global achievement tests. These results, however, do not provide a basis for selecting between the two basic strategies.

Test-retest reliability. Test-retest reliability is another traditionally accepted criterion for determining a measure's technical adequacy (Stanley, 1971). Fuchs, Deno, and Marston (1982) investigated the test-retest reliability of a simple reading performance measure. On four occasions within a period of three weeks, two third-grade oral reading in context measures were administered to 30 students in grades 1-6. Across two occasions, stability coefficients (Epstein, 1980), indicative of test-retest reliability, were +.96 and +.93 for words correct scores on the two passages. Across four administrations, the stability coefficients were +.96 and +.92 for words correct scores on both passages. Consequently, this study demonstrated test-retest reliability for two third-grade performance measures.

No study addressing the test-retest reliability of progress measures has been identified. However, within the context of repeated measurement, test-retest reliability is a not critical criterion, because measurement error is reduced when performance is sampled and summarized across many observations. Therefore, studies addressing the technical adequacy of performance and progress measures do not provide a clean basis for selecting between the strategies. Effectiveness and logistical considerations must be studied.

Effectiveness Considerations

Tindal, Fuchs, Christenson, Mirkin, and Deno (1981) contrasted the effect of performance and progress measurement and evaluation on

students' reading achievement. Approximately 80 students were assigned randomly to one of two groups: (a) a performance measurement and evaluation treatment group (PER) where reading achievement was monitored by measuring and evaluating a student's isolated word reading rate performance on random samples from a pool of words representing the student's 12-week goal; and (b) a progress measurement and evaluation treatment group (PROG) where reading achievement was monitored by measuring and evaluating a student's mastery of a series of reading vocabulary units that constituted the student's 12-week goal. The students were tested midway through, and at the end of, the 12-week intervention period on isolated and in-context reading measures. No significant differences were found between the achievement of the two groups. In a partial replication of the study, Fuchs, Deno, and Roettger (1980) found, within the context of an N = 1 study, that PROG and PER did not affect differentially a student's achievement.

Investigating the effect of measurement strategy on teacher behavior, Mirkin, Fuchs, Tindal, Christenson, and Deno (1981) found that teachers who employed progress measurement were more realistic and optimistic about their students' programs than teachers who employed performance measurement, and that they introduced fewer unnecessary program modifications. Therefore, it appears that basic measurement strategy does not affect student achievement, but that it may affect the accuracy of teachers' judgments concerning student progress.

Logistical Considerations

In the absence of technical and strongly persuasive effectiveness criteria for selecting between progress and performance measurement, it appears that teachers may prefer progress measurement. This conclusion is suggested by two pieces of evidence: (a) in a rural special education cooperative where direct and repeated measurement was adopted, teachers selected progress measurement for reading-in-context measures (Fuchs, Wesson, Tindal, Mirkin, & Deno, 1982); and (b) in a one-year evaluation of the implementation of performance measurement in reading and progress measurement in math in an ESEA Title I Program of a large city school district (Bowers, 1980), teachers agreed that the progress measurement format was more helpful in setting individual objectives, planning remedial instruction, monitoring student progress, and modifying instruction. On the basis of this evidence, one might infer that teachers preferred the progress measurement strategy.

Several logistical advantages associated with progress measurement strategies may explain teachers' preference for progress measurement using reading in context. First, progress measurement monitors mastery through a series of tasks or materials, such as a set of reading stories, a set of reading units, or a hierarchy of phonetically regular words. In this way, the procedure closely corresponds to the typical educational model wherein a child successively masters a hierarchy of skills or gains proficiency on increasingly more difficult material. Additionally, the measurement task always corresponds to the current instructional material;

therefore, the face validity of progress measurement is high. Second, the progress measurement procedure demands that the teacher determine the instructional sequence, establish criteria for mastery of each skill in the sequence, and measure performance on each skill until the mastery performance level is achieved. Consequently, the measurement procedure structures the teaching procedure, increasing the probability of improving the quality of instruction and facilitating improved student progress.

Another advantage of progress measurement is that it avoids a critical problem associated with performance measurement by eliminating the need for the teacher to select, often arbitrarily, the instructional level of the task. Each of the instructional levels that the teacher might select as the measurement task potentially is problematic. For instance, if the teacher chose the student's current instructional level, the measurement task soon might be too easy and render data insensitive to student growth. Similarly, measurement material from the age-grade appropriate level might be too difficult for a long period of time and again be an insensitive measure. Additionally, if material from the IEP goal was selected, it would be relevant only if the goal had been set appropriately. Mastery monitoring avoids this dilemma, because the measurement material continually changes as the pupil progresses.

On the other hand, performance measurement appears more feasible for two reasons. First, in performance achievement, it is unnecessary for teachers to establish a skill hierarchy and performance standard within each curriculum segment. Second, it is easier for teachers to

collect normative data on mainstream peers because they simply need to randomly select peers and administer the performance task to these peers. In progress measurement, the teacher must collect data that represent average peers' progress through the skills hierarchy; this can be a time-consuming procedure.

To assess which basic strategy is more feasible, one must consider the preceding discussion in light of his/her individual resources, needs, and values. These logistical differences between the formats should be considered in selecting between the two strategies.

How to Measure: The Selection of a Score

Once a behavior to measure and a basic measurement strategy have been selected, one must select a score to monitor, namely, correct rate, error rate, or percentage correct. The decision-making criteria for this parameter are the same regardless of whether progress or performance measurement is employed. Therefore, the following discussion does not distinguish between performance and progress measurement. It presents technical, instructional, and logistical considerations that bear on the selection of a score.

Technical Considerations

Three technical considerations relate to the selection of a score: the concurrent validity of a measurement score with respect to reading achievement tests, slope, and test-retest reliability.

Concurrent validity. Several studies provide evidence that either correct rate or percentage correct is a more valid score than error rate. Deno, Mirkin, Chiang, and Lowry (1980) found that error

rate correlations were, except for isolated exceptions, lower (r = +.39) than correlations calculated on correct rate performance (r = +.80). Furthermore, in contrast to the correlations for correct performance, many of the coefficients for incorrect performance were unreliable. Deno et al. also compared percentage correct correlations to correct and error rate correlations, and found that correlations generally were highest when percentage scores (r = +.83) were employed as the measures.

In a comparison of correct rate, error rate, and correct percentage scores, Fuchs (1981) found that correlations calculated on correct rate were higher (r = +.84) than correlations calculated on percentage correct (r = +.63) for 85% of the measures. Additionally, correlations for both correct rate and for percentage correct consistently were higher than correlations for error rate (r = +.25); and, in contrast to the correlations for correct performance, some coefficients for incorrect performance were unreliable. Therefore, within oral reading measures, it appears that correct performance scores better predict performance on standardized achievement tests.

Sensitivity to student growth. Fuchs (1981) analyzed slope as a function of how scores were calculated. In a comparison of correct rate, error rate, and percentage correct, correct rate produced steeper slopes (average = 31.2) than either error rate (average = 1.87) or percentage correct (average = 9.98). Therefore, correct rate calculated on in-context reading rendered measures on which students manifest steeper slopes and on which students have greater opportunity to register relatively small gains.

Test-retest reliability. Fuchs, Deno, and Marston (1982) investigated relative stability coefficients, indicative of a measure's test-retest reliability, as a function of type of score employed. They found that across two-day and four-day coefficients, correct rate stability coefficients were higher (+.96 and +.96) than error rate stability coefficients (+.78 and +.93). In conjunction with the above cited evidence for the greater concurrent validity and slope of correct performance calculated on oral reading measures, it appears that correct rate is technically superior to other reading measurement scores.

Effectiveness Considerations

A number of instructional reasons exist for advocating the use of rate rather than percentage. Precision teaching experts (Cohen, 1975; Haughton, 1972; Lindsley, 1971; Lovitt, 1977) argue that rate is more sensitive to behavioral change; that it provides a basis for comparing among curricula; that, in contrast to percentage, it combines speed with accuracy and imposes no performance ceiling. Further, although percentage implies a reciprocal relationship between correct and incorrect responses, this is not necessarily the case. It is possible for a student to score 90% on two days even though the student's performances on the two days were qualitatively different. The student could read more correct words as well as make more errors on one of the days, yet the score remain at 90. Correct and error rate would provide more information than percentage correct. Therefore, it appears that correct rate of oral reading from text is preferable to either error rate or correct percentage scores. However, for

effectiveness reasons, practitioners might monitor correct and error rate simultaneously in order to assess both rate and accuracy.

Logistical Considerations

In a comparison of the logistical feasibility of rate and percentage scores, rate apears to be superior. Although both scores entail the same amount of time in counting, the percentage score requires an additional division step.

Therefore, computing correct and error rate on oral reading from text appears to be the most feasible score combination. Given its technical and instructional advantages, it appears to be a sound choice for inclusion in a measurement and evaluation system.

How to Measure: The Selection of a Mastery Unit Within Progress Measurement

Certain system parameters are relevant only within one of the basic strategies. The selection of a mastery unit is a decision that applies only to progress measurement. In order to illustrate clearly why the selection of a mastery unit is a critical decision in progress measurement, it is necessary to review briefly the progress graph. The progress graph displays the rate at which a student masters a curriculum. In constructing the graph, the curriculum is segmented into mastery units--pages, stories, or clusters of stories; phonics clusters, categories, words, etc. These mastery units are placed on the vertical axis of the graph, and time (days, weeks, or months) during which the student works on the mastery units is segmented along the graph's abscissa (horizontal axis). When mastery occurs, a point is plotted at the intersection of the time and number of cumulative

units mastered. Points are plotted sequentially and connected; the resulting line depicts progress through the curriculum over time.

A problem is inherent in the progress measurement graph. Although the vertical axis is marked at equal intervals, curricula are not designed so that the mastery units sequentially plotted at these equal intervals - pages, stories, clusters of stories, even entire books - actually represent equivalent segments of the curriculum. For example, one story might be more difficult than the one it follows; while plotted as two equal units, the second story actually represents a larger unit of mastery.

While there is no available research that addresses the relative effectiveness or logistical feasibility of alternative mastery units, the selection of a mastery unit can be based on technical considerations. The selection of a very small mastery unit should increase the probability that units on the vertical axis will represent equal intervals. Pages are much more likely to be equivalent mastery units than are stories, clusters of stories, or entire books. An additional advantage in selecting a very small mastery unit is that it virtually ensures that the data will be sensitive to student change. With pages as the mastery unit, students easily can register growth. In contradistinction, if books are the unit of mastery, the measurement format might be insensitive to student improvement because the pupil would have to gain much proficiency before mastering an entire book and registering any growth. Therefore, both of these technical considerations, (a) approximating equivalent mastery units, and (b) insuring sensitivity

to student growth, support the use of a relatively small mastery unit (such as pages) within a measurement and evaluation system.

How to Measure: The Selection of a Difficulty Level Within Performance Measurement

The selection of a difficulty level is a decision relevant only to performance measurement. In performance measurement, one must select a difficulty level and maintain the same difficulty level to measure the student's proficiency changes. The measurement task is a random sample of items from a large pool of items, the difficulty level of which is constant. In performance measurement, the selection of a difficulty level is a critical decision. The following discussion covers technical, effectiveness, and logistical considerations in selecting a difficulty level.

Technical Considerations

Two relevant technical considerations for selecting a difficulty level are concurrent validity and sensitivity to student growth.

Concurrent validity. Fuchs (1981) investigated whether the criterion validity of simple performance measures was dependent on the difficulty level of the material employed. She found that within correct rate and percentage correct, all correlations between performance scores at different difficulty levels and achievement test scores were statistically significant. For error rate, all but two correlations were statistically significant. Therefore, regardless of difficulty level, performance measures demonstrated concurrent validity with respect to achievement tests.

Within measure type, an analysis of the strength of the

predictive efficiency revealed that for correct rate measures there was a weak negative association between correlation size and difficulty level. The slope of the trend line of the correlation per difficulty level, calculated with the split-middle solution (White, 1971), was an average -.0040, whereas the slope for the percentage correct was an average +.0048, indicating a weak positive relationship. For error rate, the association between correlation size and difficulty level was negative but again weak, with the correlation decreasing an average .0191 for each level increase in difficulty. Within error rate functions, there was greater variability. As indicated by total bounce, i.e. the distance between the line parallel to the trend line passing through the point farthest above the trend line and the line parallel to the trend line passing through the point farthest below the trend line (Pennypacker, Koenig, & Lindsley, 1972), variability was an average 37.5% for error rate functions, or 15.1 times greater than the mean total bounce for correct rate functions, and 3.41 times greater than the mean total bounce for percentage correct functions. As the average slope for measure types increased, the variability in the functions also increased, attenuating the significance of those increasing relationships.

Therefore, it does appear that, across difficulty levels, correct rate correlations remain relatively stable, percentage correct correlations increase slightly and demonstrate increased variability, while error rate correlations decrease and evidence relatively great variability. Given these results, the educator may prefer correct

rate measures. Within correct rate measures, the difficulty level of the material does not affect the correlation size, which represents the predictive efficiency of the measure with respect to achievement tests. Thus, difficulty level apparently is not critical when correct rate is scored. However, within error rate and percentage correct scores, the selection of a difficulty level may be critical even though it is unclear (as indicated by correlation size) which difficulty level is most adequate.

Sensitivity to student growth. Fuchs (1981) examined the relationship between the average progress per grade level and the difficulty level of the material from which students read. Employing 10 difficulty levels, she averaged scores across reading curricula and the types of measures, and found a weak, negative trend in progress as a function of passage difficulty. That is, as the passages increased in difficulty, the average progress per grade level, or sensitivity to student growth, decreased. However, within correct rate scores, which so far, on all technical bases appear preferable, the average change in score per grade level was variable, ranging from 25.92 to 38.52, with the greatest slopes in the mid range of difficulty levels.

Deno, Mirkin, Chiang, and Lowry (1980) examined the sensitivity of third and sixth grade measures. Averaged across types of scores, the mean slope for third-grade measures was 13.0; for sixth grade measures, it was 10.5. This corroborates the finding that a mid-range difficulty level maximized slope.

Mirkin (1978) also examined the relationship between difficulty level of materials and sensitivity to student growth. In this study,

each pupil was placed at three difficulty levels; independent level, where the student read between 10 and 30 words correct per minute (wpm), frustration level, where the student read between 35 and 60 wpm, and instructional level, where the student read between 50 and 75 wpm. Then, students read at each of those levels for 18 school days. Mirkin found that for correct rate performance, the average growth slope across the 18 days (calculated on the split-middle trend line) was +1.00 for the independent level, +.48 for the frustration level, and +1.03 for the instructional level. Mirkin also found that at the instructional level a greater percentage of students manifested greater growth.

These results corroborated Fuchs' (1981) and Deno, Mirkin, Chiang, and Lowry's (1980) findings, where for correct rate and percentage scores, respectively, the passages with mid-range difficulty rendered the steepest performance slopes and were indicative of heightened sensitivity to student growth. Therefore, on the basis of sensitivity and concurrent validity, the practitioner may (a) select material that represents a mid-range of difficulty for the student, and (b) employ correct rate scores when there appears to be generally greater predictive efficiency and less variability in that predictive efficiency among difficulty levels.

Effectiveness and Logistical Considerations

There are several instructional effectiveness and logistical considerations that bear on the selection of a difficulty level of material. Measuring at the instructional level provides the educator with valuable information for the purpose of instructional planning.

The teacher can observe error and fluency patterns on instructional material and plan program changes on that basis. However, this is logistically difficult to implement because a student's instructional level continually changes, and by definition the difficulty level of the performance measurement task must remain constant. Therefore, the educator must select a range of material that would encompass the student's instructional levels over a time period.

From a logistical viewpoint, it is more feasible to measure at the age-grade appropriate level; or if the age-grade appropriate level is too difficult, at some level between the instructional and age-grade appropriate levels. Instructionally, this too provides the educator with valuable information. It informs the teacher about how the student performs in relation to his/her mainstream peers.

Therefore, it appears that a solution best satisfying all the decision-making criteria may be the selection of a difficulty level as close as possible to the age-grade appropriate level without reaching a level so frustrating that it is insensitive to student growth.

How to Measure: The Selection of the Size of the Measurement Domain Within Performance Measurement

Within performance measurement, a primary decision is domain size; that is, given a difficulty level of material, what is the size of the pool from which frequent measurement tasks should be sampled? Technical, effectiveness, and logistical considerations bear on the selection of a domain size.

Technical Considerations

Two important technical characteristics of repeated measurement

are the sensitivity and the variability of the measurement. A measure's sensitivity is operationalized in this review as the slope and as the percentage of interventions resulting in student growth. Variabiltiy indicates how well performance can be predicted accurately from the data and is operationalized here as the standard error of estimate (SEE).

Sensitivity to student growth. Fuchs, Tindal, and Deno (1981) investigated the effect of domain size on the average slope of 25 students' reading performance. Three domain sizes were contrasted. First, in the most limited domain, the instructional level domain (IL), 200 words were selected randomly from each grade level. From each domain of 200 words, 25 word lists then were selected randomly. In a more comprehensive domain, words from each entire grade (GE) provided the pool from which 25 different word lists were devised. The largest domain, the across-grade domain (AG), consisted of words from all words appearing in the preprimer through fourth-grade level, with each of the 25 lists made up of words sampled from across all of these grades.

Students were measured daily on three lists: the appropriate IL domain, the appropriate GE domain, and the AG domain. Results revealed a significant difference between the average slope of performance on the IL domain (+.49) and the GE domain (+.20). There was also a reliable difference between the average performance slope on the GE domain (+.20) and on the AG domain (-.07).

The average slope decreased as the size of domain increased. The slope was steepest when the sampling procedure was limited to a 200

word subsample from the grade level in which the student was receiving instruction (IL). There was a decrease in slope when the domain represented all the grade level words; and finally, the slope fell to near zero when the domain spanned several grades.

Percentage of interventions resulting in student growth. In a related study, Tindal and Deno (1981) trained four judges to inspect, independently, 60 student graphs (three graphs for each of 20 subjects, one at each domain size--IL, GE, and AG) and determine whether each indicated intervention had a positive effect (an increase in number of words read correct per minute, or a decrease or no change in errors per minute). The percentages of positive effects were analyzed as a function of domain size (IL, GE, or AG) to assess whether the domain size influenced the measure's sensitivity to student growth. For two judges, an effect was revealed. The AG domain resulted in a statistically lower percentage of apparent effects; there was no difference in the percentages of apparent effects for the IL and GE domains. This pattern of results, however, was not replicated with the other two judges, for whom no significant differences due to domain size were revealed.

Variability. Fuchs, Tindal, and Deno (1981) also analyzed the average variability about the slope, or the SEE, as a function of the same three domain sizes. The SEE was significantly greater when performance was measured using the IL domain (.29) than when words from the GE domain were sampled (.25). No statistically significant difference in variability was found between the GE and AG domains. In the two extreme domains, the smallest (IL) and the biggest (AG), the

degree of variability was similar; there was significantly less variability in the GE domain.

Technically preferable time-series data are those with steep slopes, a high percentage of apparent effects, and minimal variability. As demonstrated above, it is unclear whether domain size affects the percentage of apparent effects; however, domain size does appear to be related to the variability and slope of the data. Clearly, the evidence suggests that a large domain, such as the AG domain, is technically inadequate. However, a lack of findings consistently favoring either the IL or GE domain precludes the selection of one of the two domains on technical grounds. The IL domain produced a steeper slope with greater variability, whereas the GE domain rendered a flatter slope with less variability. The effectiveness and logistical considerations discussed below provide some basis for selecting between the IL and GE domains.

Effectiveness Considerations

Sample size may have an impact on teachers' data utilization, and therefore on instruction. Data sampled from limited domains might provide teachers with qualitatively different information from that provided by data sampled from more extensive domains, and might lead to different program planning decisions. For example, smaller domains may be advantageous in that they provide teachers with more immediate feedback on their instructional interventions, while larger domains may be advantageous in that they provide teachers with richer data on progress towards long-term goals.

Logistical Considerations

Logistically, a larger domain is preferable because, once established as the pool from which repeated measures are drawn, it can remain intact and provide comparable data over an extended time. The final choice of domain size in a measurement system may rest ultimately on this logistical consideration. A mid-sized domain may represent the best compromise. It most probably will satisfy all requirements, rendering data with relatively low variability and with an acceptable performance slope even as it supplies an acceptably large material pool, one likely to remain relevant over a relatively long period of time. As of now, however, this judgment must be made logically rather then empirically since research on the achievement effects of domain size has not been conducted.

How to Measure: The Selection of a Measurement Frequency

Within the frequency of measurement parameter, one must determine a measurement schedule; that is, one must decide whether measurement should occur daily, twice per week, weekly, monthly, etc. Technical, instructional, and logistical considerations bear on the selection of a schedule.

Technical, Effectiveness, and Logistical Considerations

White (1971) established that in order to project a reliable performance trend, one needs to collect a minimum of seven data points. Therefore, to insure an adequate data base on which to support decisions concerning the efficacy of student programs and to avoid the unnecessary prolongment of inappropriate instructional strategies, the practitioner should collect data points as frequently

as possible. Additional evidence (Mirkin et al., 1980) suggests that daily measurement results in greater student progress than does weekly measurement. Nevertheless, teachers find daily measurement cumbersome and time consuming (Fuchs, Wesson, Tindal, Mirkin, & Deno, 1981). A compromise solution appears to be a measurement frequency of two to three times per week.

How to Measure: The Selection of a Sample Duration

Measurement is always, to some degree, time limited. Within sample duration, one must decide how long each measurement sample will last--30 seconds, one minute, three minutes, etc. Again, technical, instructional, and logistical considerations bear on the selection of a sample duration.

Technical Considerations

Several technical considerations are relevant in selecting a sample duration. Some considerations relate to analyses of group data: how sample duration affects a measure's concurrent validity with respect to achievement tests, and how it affects the variability or the standard deviation of a group's performance.

The second type of consideration relates to the reliability and validity of time-series data. Measurement theorists (Kelley, 1927; Nunnally, 1959) warn that apparently adequate technical data may have limited applicability to individual assessment. The standard error of the group may reduce substantially the relevance of group technical data in the interpretation of individual scores. Therefore, in examining the technical adequacy of formative measurement instruments used to test individual performance, it is important to investigate

measurement issues that directly relate to the reliability and validity of time-series data.

As discussed above, one characteristic of technically adequate time-series measurement instruments is that they result in low variability in the data. Reduced variability is important, because as variability between data points decreases, the reliability of the measure increases, the relative effectiveness of different instructional phases is determined more easily and quickly, and any one data point provides more information about a student's true score.

Additionally, as one judges the technical adequacy of a measurement format by investigating its influence on the variability in the data, one must examine simultaneously that format's effect on the level and slope of a student's performance. In fact, evidence suggests that characteristics of the measurement procedure itself may not only influence the variability of the data, but also affect the rate and trend of a student's performance (Ayllon, Garber, & Pisor, 1976).

Concurrent validity. Deno, Mirkin, Chiang, and Lowry (1980) examined how the duration of a curriculum-based test sample affected a measure's concurrent validity. Forty-five students in two midwestern metropolitan schools were tested on five curriculum-based measures whose concurrent validity already had been demonstrated. The students were tested on 30-second as well as on 60-second trials. Results revealed that the median correlation between the 30-second and 60-second samples was +.92, ranging between +.83 and +.97. All correlations were statistically significant ($p < .001$). Because the

60-second word recognition measures employed in this study had previously demonstrated consistently high correlations with standardized reading tests (Deno, Mirkin, Chiang, & Lowry, 1980), the study indirectly established the concurrent validity of 30-second word recognition measures with standardized reading tests.

Standard deviation. In the same study, the standard deviation of the group's performance as a function of sample duration was investigated. A standard deviation was calculated for the group scores on each 30-second and 60-second sample. Then, these standard deviations were averaged across the 30-second measures and across the 60-second measures. The mean standard deviation for the 30-second samples was 14.12; the mean standard deviation for the 60-second samples was 27.60. The discrepancy between these average values was subjected to a correlated t test, which revealed the difference to be statistically significant ($p < .001$).

The lower average standard deviation for the 30-second samples compared to the 60-second samples might be expected as a function of longer tests; with longer tests, the greater behavior range should result in a greater standard deviation. If one were to divide the 60-second standard deviation in half to account for the doubled sample duration, this value would be similar to the standard deviation of the 30-second sample, indicating that sample duration is not related to group variability of performance. Therefore, on the basis of these results, one can conclude that the 30-second duration samples, which are logistically more feasible, are as valid and reliable indices of reading proficiency as the 60-second samples.

Level of performance and sensitivity to student growth. Fuchs, Tindal, and Deno (1981) investigated the effect of sample duration on the level and slope of time-series data. Two similar second grade girls, who were seriously behind in reading, were tested daily on consonant-vowel-consonant patterned words. The experimental questions were examined through a reversal design (Hersen & Barlow, 1976), consisting of alternating daily 30-second measurement samples and daily three-minute measurement samples.

An analysis of graphed data for both students revealed that the median number of words correct per minute consistently was higher in the 30-second presentations than in the three-minute presentations. Despite this superior level of performing in the 30-second phases, the trends were relatively flat; the trends in the three-minute phases showed greater acceleration. The consistently higher median performances in the 30-second phases appeared to be related to the initial step down with each introduction of a three-minute phase.

Thus, the analyses of the relationship between the level of performance and the duration of measurement yielded conflicting results. The median levels of performance for the 30-second phases consistently were higher than the levels of performance for the three-minute phases, a comparison which demonstrated the superiority of the 30-second presentations. The analysis of the trends within the phases, however, rendered conflicting information with more strongly accelerating trends occurring in the longer presentation phases. It is possible that given longer phases for the three-minute presentations, performance under the longer measurement condition

might have surpassed performance under the 30-second presentations. Therefore, although the duration of measurement condition exhibited a consistent controlling effect, the exact nature of that effect is unclear and the superiority of one sample duration over the other was not established.

Variability of time-series data. Within the context of the above experiment (Fuchs, Tindal, & Deno, 1981), the effect of sample duration on the variability of time-series data also was investigated. The variability of each phase was summarized in two different ways. First, the total bounce was determined, and then the SEE (average variability about the slope) was calculated. Inspection of total bounce revealed that the 30-second phases were more variable than the three-minute phases. Moreover, Mann-Whitney tests on the total bounce and SEE scores revealed statistically significant differences in the variability between the 30-second and three-minute phases (two-tailed p = .037 and .043, respectively). This study revealed that the longer sample duration resulted in reduced intra-individual variability and increased reliability.

Effectiveness Considerations

Given evidence for the strong relation between time on task and student achievement, it may be that student achievement gains observed with the use of direct, repeated measurement may be a function of increased time on task during measurement activities. If this is so, then one might infer that as the measurement sample becomes longer, student achievement will improve. However, as described above, there are conflicting results concerning the relationship between the level

and slope of performance and the sample duration (Fuchs, Tindal, & Deno, 1981). This makes it difficult to establish the instructional superiority of one sample duration over another.

Logistical Considerations

The logistical consideration relevant to sample duration is readily apparent; that is, the shorter the sample duration, the more feasible the measurement system. In several of the above analyses, long and short sample durations appeared to have no effect on the technical adequacy of the measurement. Sample duration appeared to have its greatest impact on variability, with longer samples rendering more consistent, reliable performance, a desirable characteristic of time-series measurement. Instructionally, one can postulate the superiority of longer samples; yet, no empirical work supports this speculation. Logistically, the shorter durations clearly are preferable. Yet, the difference between 30-second and 60-second samples may be practically unimportant. The same cannot be said for 1 minute versus 3 minutes.

In selecting a sample duration that best fits one's needs, the designer of a measurement and evaluation system must consider the technical and instructional superiority of the longer duration and the logistical superiority of the shorter tests. Those factors then must be weighed while reviewing values and available resources.

How to Measure: The Selection of a Mastery Criterion

In terms of a given measurement system, one must establish criteria that specify when student mastery has been achieved. Within progress measurement one must determine mastery criteria for each

objective within the hierarchy of skills. Within performance measurement, one must determine criteria that specify, in terms of the long-range goal or outcome behavior, when acceptable performance has been met.

Technical Considerations

Three technical considerations bear on the selection of a mastery criterion: concurrent validity, the slope or sensitivity of the measure, and the congruency of the measure scores with respect to more widely accepted criteria.

Concurrent validity. All identified studies of the technical characteristics of different mastery criteria within curriculum-based measurement employ progress measures. Fuchs and Deno (1981) investigated whether the validity of a simple progress measure was dependent on the mastery criterion employed. The following seven mastery criteria were studied:

(1) the highest level at which, for preprimer (PP) through grade 3 books, 30-49 words per minute (wpm) with 7 or fewer errors per minute (epm), and for grade 4 through grade 6 books, 50+ wpm with 7 or fewer epm (Starlin & Starlin, 1974);

(2) 70+ wpm with 10 or fewer epm (Starlin, 1979);

(3) 100+ wpm with 0-2 epm (Haring, Liberty, & White, (undated);

(4) 95% accuracy (Beldin, 1970);

(5) 70+ wpm with 95% accuracy;

(6) for PP through grade 2 books, 50+ wpm with 85% accuracy for grade 3 through grade 6 books, 70+ wpm with 95% accuracy;

(7) for PP through grade 2 books, 50+ wpm with 85% accuracy (Powell, 1971), for grades 3 through 6 books, 70+ wpm with 95% accuracy.

These seven criteria were applied to the performance scores of 91 subjects on 10 passages in two different basal series. Fourteen sets of mastery scores were the result. Each set of mastery scores correlated significantly ($p < .001$) with scores on the Passage Comprehension and Word Identification subtests of the Woodcock Reading Mastery Tests.

However, a careful comparison among the average correlations associated with each mastery criterion revealed that at least one criterion of mastery was a differentially less effective predictor. Across its four correlations, the average correlation for Criterion 3 (+.62) was lower than any of the other average correlations by .23, and it accounted for only an average 38% of the variance in the achievement tests. Criterion 3 was the level at which a student read at 100 wpm with 0-2 errors. This criterion was the most stringent. It placed many students at easy reading levels, failing to discriminate effectively among readers with different reading skills, and resulting in lower correlations with standardized achievement tests. Criterion 1 (for PP through grade 3 books, 30-49 wpm with 7 or fewer epm, and for grades 4 through 6 books, 50+ wpm with 7 or fewer epm) consistently produced the highest correlations, with an average correlation of +.91 accounting for 86% of the variance in standardized achievement tests. The correlations produced by the remaining five criteria were similar, ranging from an average +.85 (72% of the variance accounted for in the standardized achievement tests) for

Criteria 4 and 5 to an average +.87 (76% of the variance accounted for in the standardized achievement tests) for Criteria 2 and 7.

The study demonstrated that the concurrent validity of simple progress measures with respect to standardized achievement was maintained regardless of which performance standard was employed. The findings for Criterion 3, however, were that a differential amount of variance was accounted for by the simple progress measures. Therefore, the validity of a progress measure can be affected by the mastery criterion employed. As practitioners select a mastery criterion to employ within direct and repeated measures, results of this study indicate that they might opt for correct performance rates between 30 and 70 and/or percentages between 85 and 95. Criterion 3, with a rate of 100 and an accuracy of at least 98%, was too stringent and failed to discriminate well among readers of different ability.

Sensitivity to student growth. Within progress measurement, Fuchs and Deno (1981) also investigated the relationship between the average progress per grade level and the mastery criterion employed. The range in the rate of average progress across mastery criteria was small, rendering clear interpretation of results difficult. However, the third criterion did appear to produce a differentially low rate of average progress. This leads one to infer that only the third, most stringent criterion, which also resulted in relatively poor criterion validity, differentially affected the average progress per grade. All the other criteria produced similar behavior ranges and similar sensitivity to student growth. Therefore, it appears that differential sensitivity or behavior range is not a very useful

criterion for selecting among mastery criteria in progress measurement.

Congruency. Within progress measurement, Fuchs and Deno (1981) examined the degree to which congruency between mastery level scores and teacher judgments of mastery level scores is dependent on the criterion of mastery employed, and the extent to which agreement between mastery grade scores and achievement grade scores is dependent on the mastery criterion employed. These questions supplemented the examination of the relationship between criterion validity and performance standard. Because it is theoretically possible for two measures to correlate well but agree poorly (Bradley, 1977), in selecting among mastery criteria, one might well consider congruency along with concurrent validity.

The following procedure was employed: First, two statistics were examined: (a) the percentages of students placed, by each mastery criterion, either low, high, or the same as teacher judgments of grade level and achievement test grade scores, and (b) correlated *t* tests on the difference between mastery scores and scores on criterion measures. Then, it was determined whether either of these statistics was dependent on the mastery criteria employed.

Results of these analyses demonstrated that although Criterion 1 produced the highest average correlation with achievement test scores (+.93), its levels did not agree well with levels derived from either teacher placements or test scores. On the other hand, Criterion 3, which resulted in the lowest correlations, also rendered scores that agreed poorly with both teacher placements and test scores.

Careful analysis of these data revealed that the practitioner might opt for Criterion 2, 4, 6, or 7. Criterion 2 was 70+ wpm with 10 or fewer errors. Criterion 4 was 95% accuracy. Criteria 6 and 7 employed different oral reading rates for primary (50 wpm) and intermediate (70 wpm) readers as they respectively employed 95% and 95%/85% accuracy standards. Any one of these four criteria was acceptable on three technical grounds: (a) they produced acceptable slopes; (b) they demonstrated good criterion validity (correlations between +.85 and +89); and (c) they resulted in at least 50% agreement and educationally unimportant differences (.50 grade level or less) between mastery scores and teacher judgments or test scores.

Effectiveness Considerations

Two effectiveness considerations apply to the selection of a mastery criterion. First, as discussed in "How to Measure: The Selection of a Score," for instructional planning purposes the practitioner may prefer a correct and error rate combination performance standard. Second, there has been some speculation that given more stringent criteria of mastery, retention of material might be facilitated. There is, however, no available empirical evidence to support this speculation.

Logistical Considerations

Logistical considerations provide some basis for selecting among mastery criteria. Based on the mastery criterion employed, the teacher's time commitment can differ, since computation of percentage scores requires a step that plotting raw scores directly does not. Therefore, on the basis of technical and logistical considerations, it

appears that absolute raw score criteria, with rates between 50 and 70 words correct with seven or fewer errors, are among the best that have been studied.

How to Measure: The Selection of a Procedure for Generation of Test Samples

How to generate test samples is a decision that specifies the mechanics of measurement, mechanics that must be held constant if measurement data are to be interpreted meaningfully. Generation of test samples refers to the procedure by which numerous equivalent measurement samples will be created. For selecting among alternative procedures, technical and logistical considerations apply.

Technical Considerations

There are no identified studies of technical characteristics of test generation procedures. Given the lack of empirical investigation of this issue, it may be most prudent for practitioners to accept traditional psychometric theory that advocates random selection; for, over the long run, random selection should produce equivalent samples (Hays, 1973). In contrast to other procedures, random selection has been subjected to a great deal of experimentation (Deno, Mirkin, Chiang & Lowry, 1980) and consistently has demonstrated concurrent validity with respect to achievement tests.

Logistical Considerations

While arbitrarily selecting words and/or passages is, in most circumstances, logistically more feasible than random selection procedures, there is a lack of empirical investigation concerning the effect of alternative procedures on the technical adequacy of

curriculum-based measures. Along with traditional psychometric wisdom, this lack of evidence argues for the use of random selection procedures. In response to the need for efficient random selection procedures, the University of Minnesota Institute for Research on Learning Disabilities (IRLD) has outlined relatively efficient random selection procedures for reading words in context, reading words in isolation, spelling, and written expression measures (Mirkin et al., 1981)

How to Measure: The Selection of Test Administration and Scoring Procedures

For reading, there is no array of scoring procedures. The IRLD has outlined an efficient, technically adequate, procedure for scoring oral reading (Mirkin et al., 1981). For test administration, there is scant empirical work on the effects of alternate procedures. The most persuasive arguments mitigating for or against certain test administration procedures are logistical; these logistical considerations are discussed below.

Logistical Considerations

The most feasible test administration procedures are those that involve group administration or administration either by others or by machinery. In reading, it is difficult to employ group administration because teachers must listen to individual students read. However, it is possible to decrease student-teacher interaction time and thereby to improve feasibility by having students independently tape their samples and by having teachers score those samples later. However, this does not reduce total teacher time engaged in measurement

activities because administration time equals scoring time in the domain of reading; for both procedures, the teacher spends the same amount of time listening to students read. Teachers must, therefore, determine which procedure is most feasible given their individual circumstances. For one teacher (Fuchs, Wesson, Tindal, Mirkin, & Deno, 1981), who compared taped with normal procedures in an N=1 reversal experiment, teacher time engaged in measurement was equivalent in both conditions, but the teacher preferred administering tasks during student time rather than scoring taped samples when the student was not present.

Another method for improving the efficiency of measurement in general, and test administration specifically, may be to have aides and/or cross-age peer tutors administer the measurement tasks. In a series of two single-subject experiments (Fuchs, Wesson, Tindal, Mirkin, & Deno, 1981), teachers identified and trained others (in one case, an aide; in the other, a student peer) to administer frequent measures. In both cases, total teacher time engaged in measurement (including teacher preparation time) increased (200% and 100%, respectively) during the phase in which the trainee administered the measures. However, in both cases, there was a steep decelerating trend in this phase; given the brevity of this phase (5 days and 7 days, respectively), one can speculate reasonably that given a longer phase, measurement time might have dropped to a level equal to or lower than that in which the teacher measured. It may be that as the trainees become proficient, the efficiency of having them measure would increase. Also, it may be that having trainees measure and

score would reduce dramatically teacher time because teachers then would not have to handle graphs as frequently. In the two studies, teacher satisfaction was mixed. One teacher thought that having others administer the measures was efficient; the other teacher disagreed.

Empirical work, therefore, has not demonstrated the logistical superiority of any specific administration procedures. Each designer of a measurement system must assess his/her own setting to determine which procedure is most feasible in that particular environment.

Summary

This chapter reviewed measurement procedures within the area of reading. For each decision necessary in formulating a measurement and evaluation system, alternative procedures were reviewed briefly. Then research data, speculation, and theory, as they related to each decision-making criterion, were presented. The result was a series of recommendations to designers of measurement and evaluation systems. These recommendations include:

- Measuring reading aloud from text
- Scoring and recording number of words correctly read
- Plotting (charting) either performance on equivalent passages or progress through progressively more difficult passages, depending upon individual concerns
- Using a small mastery unit for progress measurement
- Using a difficulty level approximating the student's instructional level for performance measurement
- Selecting stimuli from a mid-sized sampling domain
- Measuring two to three times per week
- Using a sample duration of one to three minutes

- Using an absolute raw score correct and incorrect criterion
- Selecting test passages randomly from the domain
- Using scoring and administration procedures that the teacher prefers

Chapter IV Outline

What to Measure: The Selection of a Task

- Technical Considerations
- Logistical Considerations

What to Measure: The Selection of a Behavior

- Technical Considerations
- Logistical Considerations

How to Measure: The Selection of a Scale

- Technical Considerations
- Effectiveness Considerations
- Logistical Considerations

How to Measure: The Selection of a Sample Duration

- Technical Considerations
- Effectiveness and Logistical Considerations

How to Measure: The Selection of Sampling Procedures

- Technical and Logistical Considerations

How to Measure: The Selection of Administration Procedures

- Technical Considerations

Summary

Chapter IV

Spelling

Gerald Tindal

In this chapter, the measurement of spelling is examined. As in the chapter on reading, discussion focuses on measurement parameters for which research data are available in the area of spelling. For each decision area, research results are presented as they apply to each decision-making criterion. Conclusions then are drawn concerning which procedures appear most appropriate for the measurement and evaluation of spelling.

What to Measure: The Selection of a Task

While it is conceded generally that spelling is of primary importance in written communication (Hammill & Noone, 1975; Wallace & Larsen, 1978), there is little agreement concerning what test behaviors validly operationalize a student's spelling competence. Since the strategy employed to measure spelling operationally defines spelling competence, attention must be given to the tasks used to quantify spelling in any test. Hildreth (1955) listed nine ways to assess a student's spelling skills: (1) writing dictated word lists; (2) writing dictated words in context; (3) detecting spelling errors in written composition and correcting the misspelled words; (4) recognizing errors in word lists; (5) completing sentences in cloze procedures; (6) writing letters; (7) copying words; (8) writing words for a timed period; (9) using a dictionary.

For the purpose of this discussion, Hildreth's categories are clustered into four sets of behavior: writing dictated word lists,

spontaneous writing, proofing, and cloze completion procedures. Both technical and logistical considerations in selecting among these are presented.

Technical Considerations

Validity. According to Cartwright (1969) and Horn (1941), dictation tests are more valid than proofing measures. Both propose that the behavior sampled in dictation tests is related more closely to what is meant by spelling. They argue that in proofing or editing the student is required only to recognize spelling errors. Proofing tests are based on the assumption that if spelling errors can be recognized, the student can avoid them in writing. An advantage of proofing is that many more words can be presented to the student (Freyberg, 1970), and responses can be scored easily by machines. The disadvantage of proofing is that since the sampled behavior does not require the writing of correctly spelled words, it would appear to be a less direct measure of spelling.

The cloze technique is another technique used to measure spelling. It requires the student to supply missing letters in a word or missing words in a sentence. Decoding and comprehension are two skills required to do the cloze task; this appears to confound the measurement of spelling with other critical skills.

The approach to testing spelling that is used commonly by teachers requires accuracy in writing words from dictation or within free writing. There is scant research to provide a basis for selecting one method over the other. Some argue that the prime objective in teaching spelling is to improve the student's spelling

accuracy in everyday writing (Freyberg, 1970; Rowell, 1975), and therefore, that the content validity of free writing spelling is greater.

Nevertheless, dictation tests are the most frequently used procedures for measuring spelling in schools, and according to Cartwright (1969) and Horn (1941), they are the most valid. The dictation of word lists has been found to be more successful than presenting words in sentences or paragraph form, because word selection is easier (Horn, 1944, 1954).

To identify simple procedures for teachers to use in measuring spelling, a series of three studies was conducted by Deno, Mirkin, Lowry, and Kuehnle (1980). In this research, several approaches to measuring students' spelling were examined. In the three studies, students were selected randomly from two Minneapolis schools.

Student performance in spelling from dictation and in written compositions was correlated with performance on the Test of Written Spelling (Larsen & Hammill, 1976) in the first study. The dictated word lists consisted of randomly selected words from Basic Elementary Reading Vocabularies (Harris & Jacobson, 1972). The sample of written expression was a 5-minute composition based on a picture stimulus. High correlations were obtained between performance on the dictated word lists and performance on the standardized achievement test (+.85 to +.96). Moderately high correlations were obtained between spelling within the written expression sample and the standardized achievement test (+.70), and between spelling from the dictated lists and spelling within written expression (+.61 to +.92).

In the second study, correlations between four word lists sampled from different grade levels of the Harris and Jacobson word list (1972) were correlated with the Peabody Individual Achievement Test (Dunn & Markwardt, 1970). Again, the correlations were high, ranging from +.81 to +.94. In general, list difficulty had little effect on the validity of the dictated word list measure. Structuring the word lists from easier to more difficult words produced slightly lower correlations.

In the third study, three random selections of words from the Harris and Jacobson list (1972) and one cumulative list of words selected arbitrarily from Level 9 of Ginn 720 were dictated to students. Spelling performance on these word lists was correlated with performance on the spelling section of the Stanford Achievement Test, Primary III (Madden, Gardner, Rudman, Karlsen, & Merwin, 1978). Again, high correlations were obtained (+.80 to +.89). The words arbitrarily selected from the vocabulary list in Level 9 of Ginn 720 also discriminated well among spelling proficiencies (r = +.89).

The results of these three studies indicate that student performance in spelling words selected from basal readers correlates highly with performance on standardized spelling achievement tests. In addition, performance across the various word lists in the three studies intercorrelated highly (whether the words were randomly selected, nonrandomly selected, or ordered in difficulty). These results indicate that regardless of the type of word list used, spelling from dictation validly discriminates among students of differing spelling proficiency.

Sensitivity to student growth. For a measure of spelling to be useful to a teacher, it must be sensitive to increases in student performance throughout the period of teaching. Sensitivity is related to the opportunity for the student to respond, because the opportunity to respond will have a direct impact on the quantity of behavior that can be measured. If there is very little behavior to measure, the scale will be abbreviated with less opportunity to measure difference. In a system with considerable opportunity to respond, more behavior can be measured; as a consequence, changes can be registered. The cloze and proofing techniques provide a large number of opportunities to respond; however, as discussed in the next section, with cloze or proofing tasks, the number of items actually presented by a teacher will be limited by the logistics of preparation time. Word lists also may be expanded easily to include many opportunities to respond. Given individual differences in the number of words generated in writing samples, it is unclear how much opportunity to respond might be available in free writing. Opportunity to respond in spelling from dictation also can be increased by using scoring procedures that involve units smaller than whole words. The sensitivity of doing so will be described later under "Selection of a Behavior."

While it appears that, from a technical standpoint, using dictated word lists is a valid and sensitive measure of spelling, it is important to consider the logistical and effectiveness issues with respect to measuring different behaviors. Unfortunately, there is no available research concerning effectiveness (i.e., the impact on student performance of measuring one behavior rather than another).

The remaining discussion is confined to logistical considerations.

Logistical Considerations

Logistical considerations refer to the ease with which a measurement procedure can be organized and implemented by teachers. Of the various procedures listed earlier in this unit (dictated spelling tests, cloze, proofing, and free writing), it appears that the amount of preparation time necessary to develop the dictated spelling lists may be less than for either cloze or proofing tests. The cloze and proofing procedures demand the development and duplication of written material for student use in testing. The development of single spelling lists to be used in dictation by the teacher avoids most of this preparation time. Free writing, on the other hand, involves the least amount of preparation time; however, as previously noted, spelling in free writing fails to correlate as well with standardized measures. Also, in contrast to the other test tasks, item difficulty is impossible to control in free writing tasks. This may account for the finding that spelling accuracy in free writing correlates less well with other achievement measures.

Another logistical consideration in selecting a measurement task relates to the need to be able easily to generate equivalent tests. Measurement for formative evaluation requires frequent measurement of student performance to determine the effectiveness of instructional changes. Many tests must be created and it is imperative that changes in the measurement system itself do not occur. With the use of cloze or proofing procedures, the development of parallel forms to be used 3-5 times a week is difficult and time consuming for the teacher.

Consequently, from both a technical and logistical perspective, it appears that writing words from dictated lists is the best choice for measuring spelling. It is valid with respect to standardized achievement tests. It potentially renders data that are sensitive to student growth. Further, the difficulty of the test can be manipulated easily, and parallel forms are generated readily.

What to Measure: The Selection of a Behavior

Given the decision to measure students' written spelling through list dictation, it then becomes necessary to determine what response unit should be scored. Two alternative response units are (a) writing letters in correct sequence within words, and (b) writing correctly spelled words. A discussion of the technical and logistical advantages of each alternative is presented here.

Technical Considerations

Concurrent validity. A comparison of the concurrent validities of words and letter sequences with respect to raw scores on three different published, standardized achievement tests indicates that correct words per minute and letters in correct sequence per minute scores produce similar correlations. The correlation between incorrect letter sequences or words spelled incorrectly and the number correct on the standardized achievement measures is less strong (Deno, Mirkin, Lowry, & Kuehnle, 1980). In the three studies from which these results were obtained, the informal measures were administered for a fixed period of time, ranging from one to three minutes. In contrast, none of the criterion measures was timed. The resulting high correlations indicate that students are differentiated in a

similar manner in tests of both power and speed. Low performers spell less proficiently than good spellers whether or not the measure is timed.

In the first study, a comparison was made among the correlations between words correct per minute scores and correct letter sequences per minute scores with respect to the Test of Written Spelling (Larsen & Hammill, 1976). The relation between words spelled correct and performance on this test was very high, with correlations ranging from +.83 to +.96. The correlations ranged from +.80 to +.94 when correct letter sequences was the response unit, and they remained very high when the sample included LD students only (+.89 to +.97) and regular students only (+.90 to +.95). Further, the intercorrelation of performance on these word lists showed that correct letter sequences correlates very highly with words spelled correctly, both within (+.94 to +.95) and across (+.82 to +.97) various lists.

In a second study, the Peabody Individual Achievement Test (PIAT; Dunn & Markwardt, 1970) served as the criterion measure. In this test, the student actually does not spell out words, as in the Test of Written Spelling, but rather is presented four choices (words) and directed to choose the correctly spelled word. The formative measures again consisted of three-minute dictated spelling tests. High correlations resulted for all the lists with the PIAT for both number of correct letter sequences (+.80 to +.90) and number of correct words (+.83 to +.94). Again, the correlations remained quite high when the sample included regular students only (+.81 to +.93). However, with the LD students, the correlations were considerably lower (+.29 to

+.95). This may have been a function of the LD students' restricted range of scores on the PIAT, or the difference in the unit of response between the two types of tests. In the PIAT, the student need only select the correctly spelled word from four choices, while the dictated formative measures required actual spelling. The format in the PIAT may not discriminate among students of varying spelling proficiency who have been pre-selected from the low end of the distribution.

The spelling section of the Stanford Achievement Test, Primary III (Madden et al., 1978) was used as the criterion measure in the third study. The response format for this test is similar to that used by the PIAT. Each item consists of four words, three of which are spelled correctly, and one of which is spelled incorrectly. The student is directed to identify the misspelled word. The format for the formative measure again utilized dictated word lists. Consistent with the previous findings, high to very high correlations were found among the various lists and the criterion measure. The correlations using correct letter sequences as the response unit ranged from +.80 to +.86 with all students combined. For the LD sample, the range was +.78 to +.82, whereas for the regular sample the range was +.76 to +.86. Using words spelled correct, very similar correlations were found. For the combined sample, correlations ranged from +.83 to +.89; for the LD sample the range was +.80 to +.84; for the regular sample, the correlations ranged from a low of +.80 to a high of +.89.

Concurrent validity indicates the degree to which performance on a measure relates to or predicts performance on other more

traditionally accepted measures. Because word and letter correct scores on the dictated word lists correlate so strongly with performance on the standardized achievement measures, it is possible to predict with a fair degree of accuracy a student's performance on a standardized test by knowing how a student performs on the informal measure. Because standardized tests have few parallel forms and are so time consuming to administer, a viable alternative is frequent measurement on this informal measure. Therefore, it appears useful to employ words or letters correct performance on the informal measures to assess student growth on a continuous basis; one can expect growth on the simple measures to correspond to growth on a psychometrically sound, traditionally accepted achievement test.

Concurrent validity with respect to classification also was examined. Two types of concurrent validity with respect to classification are predicting a student's (a) special education status, and (b) grade level membership. Deno, Mirkin, Lowry, and Kuehnle (1980) investigated these two types of predictive validity for the spelling measures. First, the authors compared the performance of students receiving LD services to that of students in the regular classroom. Clear differences were evident in the performances of these two groups on the dictated word lists. In eight of nine comparisons, the performances of the two groups were significantly different on the dictated word tests. This was true whether their performance was scored in terms of correct letter sequences or words spelled correctly. The raw score difference between the average performance of the LD and regular students was quite large;

performance of regular students exceeded the LD students by a ratio of 2 to 1.

A cross sectional analysis by grade level revealed increases in performance for grades 2-5; increases were apparent for both the numbers of correct letter sequences and words, with a leveling off between fifth and sixth grades. Therefore, it appears that both words correct and letters in correct sequence discriminate among students' programs and grade placements. On the basis of the validity criteria examined, one can support the selection of either words or letter sequences as the behavior to measure, but one cannot select easily between these two response units.

Interscorer reliability. Another technical consideration in the identification of what to measure is whether the unit of behavior can be identified consistently by different observers; that is, given the same sample, will different scorers obtain the same score? A measurement system must be defined and implemented objectively so that others may obtain the same results. Deno, Marston, Mirkin, Lowry, Sindelar, and Jenkins (1982) investigated the reliability of scoring correct letter sequences as well as words correct. Regardless of the scoring system, the interscorer reliability was high, ranging from +.94 to +.99 agreement, with the majority of agreements at 99%. Thus, both response units qualify on the basis of scorer objectivity.

Sensitivity to student growth. As discussed above, an important technical characteristic of a continuous evaluation system is its sensitivity to student growth. In an investigation of students' improvement in spelling from fall to spring, Deno, Marston, Mirkin,

Lowry, Sindelar, and Jenkins (1982) tested 566 students at the beginning and end of the school year. At all grade levels, the average within-student increase was statistically significant. There was nearly equal improvement across grades 1-6 from fall to spring, with the increase ranging 4.3 to 5.7 words spelled correct. The increase in correct letter sequences ranged from 29.5 to 37.7 across the six grades. Therefore, increases in spelling skill would be more apparent when letter sequences are graphed than when words correct are graphed.

The percentage of growth was greatest in the first grade (436% in the number of correctly spelled words and 384% in the number of correct letter sequences), probably because of the lower initial performance level of the first graders. The remaining grades showed a range of growth from 24% to 176% in the number of words spelled correctly, and a range of growth from 39% to 88% in the number of correct letter sequences. For the total sample, the percentage of growth was well over 100. When each grade was blocked by quartiles (25th percentile, 50th percentile, and 75th percentile), there was evidence that all students improved, regardless of the quartile in which they fell.

On the basis of the preceding discussion of technical considerations, the following conclusions are warranted: (a) correct scores are preferable to error scores because they demonstrate stronger relationships with standardized tests, with special education status, and with grade placement; (b) words correct or letter sequences correct are acceptable in terms of their strong interscorer

reliability; and (c) letter sequences correct are preferable to words correct due to more sensitive data rendered when letters in correct sequence are scored. However, before determining the score to be employed for spelling, logistical issues need to be considered.

Logistical Considerations

In the previous section, technical considerations for using letter sequences and words correctly written were discussed. The conclusion was that, on technical grounds, correct letter sequences was the most appropriate behavior to measure. However, important logistical differences exist between these two behaviors. Although the use of correct letter sequences is a valid and sensitive measure of spelling, it is also time consuming.

In a rural educational cooperative in Minnesota, it took teachers a median 4 minutes 26 seconds to score students' spelling using letters in correct sequence as the unit of measurement, in contrast to 2 minutes 32 seconds required to score the number of words spelled correctly (Fuchs, Wesson, Tindal, Mirkin, & Deno, 1981). Because of this large difference in efficiency, it may be that words correct represents an adequate score for some students, specifically more proficient spellers for whom the problem of a floor effect and of an insensitive measure is less likely. Therefore, the decision to use letter sequences must depend upon the student's level of performance. For extremely low functioning students, the proper score appears to be letter sequences. For students with some minimal level of spelling proficiency, counting the number of words written correctly may be an appropriate and efficient behavior to measure.

How to Measure: The Selection of a Scale

In each study described above, a fixed testing time was used, ranging from one to three minutes. Various scoring procedures then were implemented, including the absolute number of correct and incorrect responses, the rate correct and incorrect per minute, and the percentage correct. The range of correlations between number, rate, and percentage correct was very high (+.89 to +.97) and all three correlated highly with the Test of Written Spelling (Larsen & Hammill, 1976), ranging from +.91 to +.97. This finding, however, is predicated on the constant testing procedure, which involved a fixed time interval.

Other viable testing procedures that could be investigated include the use of no time limits or the use of a fixed number of words with no time limits. In such a testing format, there may be considerable change in the student's performance, using either number, percentage, or rate correct (assuming the testing time was measured but not pre-determined). Similarly, the resulting correlations between the scales may change. Although there is no a priori reason to predict the direction of the change, the fact that high correlations resulted between each of the three scaling formats and an untimed achievement test indicates the change may not be great. Students are rank ordered in a similar way when the testing involves either speed or power. Therefore, from a technical standpoint, teachers might select the scale they prefer; logistically, teachers may prefer a fixed, relatively short, timed test.

How to Measure: The Selection of a Measurement Frequency

One of the most critical components of a formative evaluation system is the frequency with which the behavior is measured. Data must be collected frequently in order to provide accurate and timely information regarding student performance, information that can be used to evaluate programs concurrent with their implementation. In this section, measuring spelling two to three times per week is recommended. Support for this recommendation is provided by technical, effectiveness, and logistical considerations.

Technical Considerations

Technical support for measuring two to three times per week is derived from empirical work on the use of program evaluation data. If rate of behavior change is to be evaluated, program decisions must be based on a minimum of seven data points in order to calculate the slope of a student's performance. White (1977) found that with nine days of data, performance could be predicted into the following week with 64% accuracy. With eleven days of data, the accuracy of this prediction rose to 85%.

Clearly, then, to obtain a stable and reliable index of the rate at which the student is improving over time, more data points are better. With measurement occurring twice per week, a minimum of three and one-half weeks of data is necessary before a valid assessment may be obtained; at best, with 36 weeks of school, 12 instructional changes or strategies could be attempted throughout the year. Measurement of spelling three times a week would allow for a more reliable and responsive system, providing a broader range and/or

greater frequency of changes.

Effectiveness Considerations

Available empirical data tend to support a daily measurement system. Mirkin et al. (1980) found that when students were measured daily in spelling, they improved more than when their performance was measured on a weekly basis. Although the focus of this study was on how students' performance data were utilized to make program changes, there was a clear indication that frequent measurement, even without teacher inspection and use of data, could be effective in improving performance. However, daily measurement raises logistical issues.

Logistical Considerations

Although it appears that the more data available, the better for evaluating student change, there is a point at which the cost of collecting these data may outweigh the benefits. In the case of spelling, this is a very important issue because the scoring of spelling can be cumbersome. Therefore, efficiency warrants measurement of performance on only two or three days each week.

How to Measure: The Selection of a Sample Duration

Length of test has an influence on both the technical adequacy and the practicability of the evaluation system. The test must be long enough to provide a representative sample of behavior, yet short enough so that time spent preparing for and executing the measurement task does not become cumbersome. As demonstrated below, technical, instructional, and logistical considerations support the use of a two-minute spelling sample.

Technical Considerations

Validity. In studies by Deno, Mirkin, Lowry, and Kuehnle (1980), examiners dictated words from a list for three minutes, and recorded the students' performances after one, after two, and after three minutes. The intercorrelations for these three test durations revealed a high relationship among performances for all three durations for both letter sequences and words (+.79 to +.93). Approximately 50% of all correlations were above +.90. The correlations were slightly higher when letters in correct sequence was the unit of measurement. Therefore, it appears that consistent information on a student's spelling skill will be provided by any of the three test durations.

Additionally, the correlation of performance on the spelling lists of different durations with performance on both the PIAT and the Stanford spelling test ranged from +.67 to +.92. There was very little difference in the strength of the correlations between the word lists and the two standardized tests for each test duration; that is, one minute samples and three minute samples yielded similar correlations with both standardized achievement measures. The correlations between correct words and the standardized tests were slightly higher than those between letters in correct sequence and the standardized tests.

Sensitivity to student growth. Another relevant technical consideration in selecting test duration is the sensitivity of the resulting data to student growth. At present, no research is available on the degree to which sample duration influences the

sensitivity of the spelling measures. However, it has been found that many special education students have poor spelling skills and, given a short spelling test, they typically write few correct words. This is problematic in light of the need to provide ample opportunity for the behavior to occur if the measure is to be sensitive to student growth. With a two or three minute sample, a special student with poor skills is less likely to show a floor effect in which the student emits no response. Consequently, while either sample duration demonstrates validity with respect to standardized tests, consideration of the sensitivity of the measurement to student growth provides support for a relatively long sample. However, effectiveness and logistical issues have yet to be considered.

Effectiveness and Logistical Considerations

In a continuous evaluation system, both instructional effectiveness and logistical considerations clearly bear on the issue of sample duration. A longer sample duration is supported by literature demonstrating that academic engaged time is correlated highly with student achievement (cf. Graden, Thurlow, & Ysseldyke, 1982; Greenwood, Delquadri, Stanley, Terry, & Hall, 1981). Spelling words during a test is a form of student engagement and an increase of one to three minutes of active academic responding two or three times per week may increase student growth. However, the teacher must be able to use the measure efficiently. Logistically, a one minute sample is less time consuming for teachers than a two-minute or three-minute sample.

On the one hand, then, instructional and technical considerations

support long samples. On the other hand, efficiency considerations support short samples. A compromise solution appears to be the use of a two-minute sample.

How to Measure: The Selection of Sampling Procedures

Two major sampling questions are: (a) what should be the domain from which the words are drawn, and (b) what procedures should be employed to generate alternate test forms? Relevant considerations in answering these questions are both technical and logistical. For the first question, the concern is how the domain size affects the sensitivity of the measurement system to student achievement. With respect to the second question, the concern is how the sampling procedures affect the concurrent validity of alternate forms.

Technical and Logistical Considerations

With respect to the relation between domain size and the sensitivity of the measurement, no experimental contrasts of sampling domains have been conducted. However, there is empirical evidence that more than one domain size renders scores sensitive to changes in spelling performance over time. In an evaluation of teachers' use of spelling data for making program changes, Mirkin, Deno, Tindal, and Kuehnle (1980) found that treatment effects were evident both when a within-grade-level list of words was used and when a list of words from across several grade levels was used.

Further, in a study involving the measurement of students in the fall and again in the spring (Deno, Marston, Mirkin, Lowry, Sindelar, & Jenkins, 1982), there were significant increases in performance for both the number of words spelled correctly and letters in correct

sequence with a sample of words from a cumulative preprimer-grade 3 list. Evidence suggests, then, that by sampling from material around the instructional grade level, one can obtain a sensitive measure of spelling. Apparently, it is not essential that the measurement domain be limited to current instructional words, but rather it may include items within one to two years of the student's instructional level. With this domain size, one must determine the sampling procedure or the method by which alternate tests will be sampled.

Two studies have addressed how procedures for generating test samples affect the criterion validity of curriculum-based spelling measures (Deno, Mirkin, Lowry, & Kuehnle, 1980). Both of these studies examined alternate procedures for generating spelling test words. In the first investigation, correlations between achievement test scores and scores on either random lists of words or lists of words arbitrarily selected from the backs of books were approximately equivalent. This finding indicates that these test sample generation procedures do not affect the criterion validity of curriculum-based measures. Nevertheless, within the context of frequent measurement, random lists theoretically should represent more equivalent samples over the long run (Hays, 1973), and thereby should enhance the validity of the measure and the interpretability of the performance data.

In a second study, a random selection procedure was contrasted to one in which words were ordered from easy to difficult, progressing from preprimer through sixth grade level. Examination of correlations with achievement test scores revealed that both procedures yielded

significant correlations; however, the random selection procedure resulted in higher criterion validity coefficients than the ordered lists.

Within dictated spelling test measures, it therefore appears that both a random generation procedure and a procedure in which a teacher arbitrarily selects words from those in the backs of books demonstrate criterion validity and are equally good predictors of achievement test scores. While this was true in a cross-sectional correlational study, it is unknown whether the results would be demonstrated in a longitudinal study. Additionally, a procedure that generates ordered lists of words appears to lack concurrent validity. These studies provide limited information concerning the impact of alternate sampling procedures; yet, given the lack of further empirical investigation, it may be most prudent for practitioners to accept random selection.

How to Measure: The Selection of Administration Procedures

Research on the effect of administration procedures on the measurement of spelling is scant. According to Horn (1941), "the most valid and economical test is the modified sentence recall form, in which the tester pronounces each word, uses it in an oral sentence, and pronounces it again. The word then is written by the students" (p. 1179). Although not the focus of research, several alternative procedures have been used in various studies conducted by the Institute for Research on Learning Disabilities. In the absence of direct empirical data, review of these procedures provides a logical basis for recommending 15-second, paced dictation. With this

procedure, a new word is presented immediately after the student has finished spelling the previous word; however, if 15 seconds have elapsed and the student has not finished spelling the word, a new word is presented. In the directions provided prior to testing, these procedures are explained to the student who is asked to keep up with the words presented by the examiner.

Technical Considerations

Support for the 15-second paced dictation administration primarily is technical; namely, its potential effect on the sensitivity of the measurement system. Without the use of rolling dictation the possibility of a floor effect or zero response level would increase. If a new word was presented only after the student had completed the previous one, the amount of behavior sampled could be limited. For students completing the task at a laboriously slow pace or for students who simply cannot spell the word but continue to work on it, few words would be attempted, and student growth over time would appear minimal. Paced dictation attempts to reduce this possibility and to increase the probability that the measurement will be sensitive to student achievement.

Summary

The discussion in this chapter supports the following conclusions:

- The behavior measurement task should be writing words dictated from lists
- For low functioning students, the preferable behavior to measure is correct letter sequences, and for more proficient students the preferable unit is words spelled correctly
- Measurement should occur at least two times per week

- Each test should last for two minutes
- Test items should be sampled randomly from the measurement domain
- 15-second paced dictation is an acceptable administration procedure

Chapter V Outline

What to Measure: The Selection of a Behavior

- Technical Considerations

How to Measure: The Selection of a Scoring Procedure

- Technical Considerations
- Logistical Considerations

How to Measure: The Selection of a Stimulus Format

- Technical Considerations
- Logistical Considerations

How to Measure: The Selection of a Measurement Domain

- Effectiveness Considerations
- Logistical Considerations

How to Measure: The Selection of a Measurement Frequency

Summary

Chapter V

Written Expression

Doug Marston

Chapter V is an examination of the measurement of written expression. As with reading and spelling, the questions to be answered are "what to measure" and "how to measure" when formatively evaluating written expression. Three measurement parameters are not discussed in this chapter because they are not applicable to the measurement of written expression: the selection of a sampling procedure, a basic strategy, and a mastery criterion. For each decision area discussed, available technical, effectiveness, and logistical research data are presented and recommendations for measurement procedures are made.

What to Measure: The Selection of a Behavior

Several alternative measures of written expression may be incorporated into a continuous evaluation system. In selecting the most appropriate measure, the primary considerations are technical; that is, what behavior or behaviors can be measured reliably and validly to reflect growth in writing connected discourse.

Technical Considerations

Deno, Marston, and Mirkin (1982) identified six potential measures of written composition: Mean T-unit Length (Hunt, 1966); "Mature Words" (Deno, Marston, & Mirkin, 1982), Total Words Written (Myklebust, 1965), Large Words, Words Spelled Correctly, and Correct Letter Sequences (White & Haring, 1980). The technical adequacy of these measures was determined by examining criterion validity,

discriminative validity, reliability, and sensitivity to student growth.

Criterion validity. The study of criterion validity (Deno, Marston, & Mirkin, 1982) focused on the six potential measures of written expression and their correlations with standardized achievement measures. Using the Test of Written Language (Hammill & Larsen, 1978), the Word Usage subtest from the Stanford Achievement Test (Madden et al., 1978), and the Developmental Sentence Scoring System (Lee & Canter, 1971) as criterion measures, it was found that Mature Words, Total Words Written, Words Spelled Correctly, and Correct Letter Sequences correlated significantly (+.70 or higher) with the criterion measures of written expression. The magnitude of the coefficients indicated that these simple procedures are valid measures of written expression.

Discriminative validity. Deno, Marston, and Mirkin (1982) also examined whether simple, direct measures of written expression discriminated students receiving learning disability services and children enrolled in regular classes. Reasoning that students with learning problems would do much poorer on valid written expression measures than regular students, the authors examined the group differences. Results indicated that, within grades 3, 4, 5, and 6, Total Words Written, Mature Words, Words Spelled Correctly, and Writing Correct Letter Sequences significantly differentiated these two groups. These significant differences provide a second source of validation for the simple, direct measures of written expression.

Reliability. Test-retest reliability indicates how well a

measure provides consistent results across time. For example, if Total Words Written demonstrates high test-retest reliability, then one might expect that the student's relative performance in writing a composition today will be equivalent to his/her relative performance in composition writing tomorrow. When relative test-retest scores are inconsistent, the teacher is more likely to make incorrect decisions regarding student achievement.

Parallel-form reliability also is important to the teacher implementing a formative evaluation system for written expression. Since this assessment approach is based on repeated measurements, the teacher needs many different forms for sampling written expression. It is imperative that these parallel forms provide equivalent results. For example, one should expect that the relative performance in the Total Number of Words Written for one story starter is equivalent to the relative performance in Total Words Written for a different story starter if the two writing formats have high parallel-form reliability.

Marston and Deno (1981) explored the test-retest and parallel-form reliability of simple direct measures of written expression. Correlation coefficients above +.80, indicative of reliability, were found for Total Words Written, Mature Words, Words Spelled Correctly, and Letters Written in Correct Sequence. Marston (1982), however, found low correlations for single writing samples and cautioned that one sample of written expression may not be entirely reliable. Aggregating three writing samples and using the mean resulted in acceptable reliability coefficients (above +.80).

Sensitivity to student growth. While it is important that the measure of written expression employed in direct and repeated evaluation of student progress be useful in discriminating groups, it is even more important that the measure be useful in discriminating growth within the same individual. Formative evaluation involves frequent decisions regarding program effectiveness. To make these decisions, data on performance changes of relatively small magnitude over short time periods are necessary. Standardized achievement tests generally are not constructed for these purposes. The factor considered here is what unit of behavior is most likely to be sensitive to growth in proficiency in writing connected discourse.

To answer this question, the actual growth (increments in performance) on the simple measurement procedures was investigated by Deno, Marston, and Mirkin (1982). They examined the scores of 130 elementary students who wrote 3-minute compositions in response to story starters. The mean performance for each grade level indicated that the range of performance for Grades 3-6 was greatest for Correctly Written Letter Sequences. The range was moderate for Total Words Written and Words Spelled Correctly, and small for Mature Words. As may be seen in Figure 3, when the data for regular students were graphed for visual analysis, progress was most apparent for Correct Letter Sequences, where performance increased from 153.5 to 270.0 across four grade levels. Mature Words, on the other hand, ranged only from 9.6 to 14.6. Similarly, for LD students the scores for Correct Letter Sequences ranged from 18.3 to 164.8, whereas those for Mature Words ranged from 1.7 to 9.4. This phenomenon is significant

given the central role of graphing in formative evaluation procedures. For example, if the teacher is monitoring student progress on Mature Words, it will be difficult to ascertain visually on a graph whether growth has occurred. Yet a one-word increase on this measure may signal marked improvement. The larger ranges of Correct Letter Sequences, and to a lesser degree, Total Words Written, permit one more easily to discern growth on a chart.

Insert Figure 3 about here

While the data across grade levels suggest that Correct Letter Sequence measures are more sensitive to growth, they do not bear on the issue of growth within individuals at a given grade level. To examine this question, Deno, Marston, Mirkin, Lowry, Sindelar, and Jenkins (1982) analyzed the gains made by 566 elementary students on Total Words Written, Words Spelled Correctly, and Correctly Written Letter Sequences. Students were measured twice within the same academic year, once in the fall and again in the spring, with approximately six months intervening. Paired _t_ tests were applied to students' fall and spring test performance within each grade (1-6) to ascertain the statistical significance of students' gains. Sixteen of eighteen _t_ values were significant at or below the .01 probability level. Only performance on Total Words Written at sixth grade failed to render a significant increase from fall to spring. Therefore, the _t_ test analysis supports the notion that the simple direct measures of written expression are sensitive to growth within grade levels, with

the most dramatic effects at Grades 1 and 2, Grade 1: $t(91) = 9.72$, $p = .001$; Grade 2: $t(84) = 7.44$, $p = .001$. The same pattern of results was obtained when percentage growth was examined rather than absolute gain.

In summary, adequate evidence exists for the technical adequacy of three direct measures of written expression, Total Words Written, Words Spelled Correctly, and Correctly Written Letter Sequences, with Letter Sequences perhaps somewhat better. Data suggest that the practitioner may choose confidently among these three measures on the basis of their demonstrated validity, reliability, and sensitivity to student growth.

How to Measure: The Selection of a Scoring Procedure

The behaviors examined by Deno, Marston, and Mirkin (1982) differ with respect to scoring procedures required to generate numerical data. The scoring approaches vary from simply counting the number of words written in a composition (Total Words Written) to analyzing and counting the number of letters written correctly (Letters in Correct Sequence). The technical and logistical problems of these scoring procedures are reviewed here.

Technical Considerations

The first factor in determining the technical adequacy of the scoring procedure is the degree of interscorer agreement that is obtained when written samples are analyzed. If the scoring procedure for a measure of written expression was technically adequate, then one would expect that different judges would arrive at the same scores when analyzing the composition. Videen, Deno, and Marston (1982)

examined this question by having four undergraduates in special education score the same set of 20 compositions. To determine the extent to which the judges agreed, a correlation coefficient was computed for each pairwise judge combination. The average Pearson correlations among judges were: for Total Words Written, +.98; for Words Spelled Correctly, +.98; and for Correct Letter Sequences, +.99. These high correlations provide evidence for interjudge agreement and technical adequacy of all of the scoring procedures.

Logistical Considerations

When considering the logistics of the scoring procedures, time factors become crucial; the various direct measures have differential time demands. In an analysis of the efficiency of scoring procedures (Videen et al., 1982) Total Words Written appeared to take the least amount of time to score. The Total Words scoring procedure, averaged over 20, 40, and 60-word compositions, took 25% less time than counting Words Spelled Correctly and 83% less time than scoring Correct Letter Sequences. These data suggest that the preferred practice with respect to logistics is Total Words Written or Words Spelled Correctly.

How to Measure: The Selection of a Stimulus Format

Several approaches to collecting written compositions from elementary students are available to the teacher. Myklebust (1965) suggested that a picture stimulus is helpful in eliciting writing samples from children. With this approach, the teacher shows students a picture depicting some activity and asks them to compose a story about the scene. After students are given several minutes to think

about their story, they generate written compositions. Deno, Mirkin, and Marston (1980) explored two other approaches to elicit writing samples. The first method employed a story starter, where the child was read a sentence and then asked to write a short story. For example, students might be asked to write a story that began with "One night I went outside when it was very dark." The second method involved the use of a topic sentence. The topic sentence directs the student to simply write about a prescribed topic. For example, a student is asked to "write about summer vacation." Relevant considerations for selecting between these stimulus formats are technical and logistical.

Technical Considerations

Validity. The relative validity of the three stimulus formats was determined by comparing correlations between student performance on the Test of Written Language (Larsen & Hammill, 1976) and Developmental Sentence Scoring (Lee & Canter, 1971) and the direct measures based on the various writing formats (Deno, Mirkin, & Marston, 1980). Correlations were high and similar for performance based on all three stimulus procedures: differences among the approaches were unimportant. These data suggest that the teachers may have confidence in the validity of using verbal or pictorial stimuli to generate writing samples.

Reliability. One approach to determining the reliability of the various stimulus formats is to analyze the internal consistency of each methodology. The performance of students was compared at the end of minutes 1, 2, 3, 4, and 5 (Deno, Mirkin, & Marston, 1980). Average

intercorrelations were +.87 for the Story Starter, +.94 for Topic Sentence, and +.92 for the Picture Stimulus. The coefficients indicate that a high degree of internal consistency reliability exists for the three formats.

Logistical Considerations

Myklebust (1965) has made several practical recommendations regarding the use of a picture stimulus. He suggested that the picture involve activity, color, and several characters. Additional practical considerations include picture size and cost. (If the written measure is to be group administered, the teacher will find larger pictures advantageous.) Some commercial educational products containing pictures are available (Dunn & Smith, 1967), but quite expensive; the resourceful teacher can find picture stimuli in books, magazines, and newspapers.

Given these considerations, Story Starters and Topic Sentences probably are more feasible than picture stimuli. They are less expensive to produce and can be incorporated more easily into the response form. Deno, Marston, and Mirkin (1982) used Story Starters printed at the top of lined paper. This format allowed students not only to listen but also to look at the stimulus. Therefore, it appears that Story Starters or Topic Sentences would be less time consuming and expensive than picture stimuli.

How to Measure: The Selection of a Measurement Duration

Deno, Mirkin, and Marston (1980) examined the written compositions of 51 elementary students attending regular classrooms and 31 children enrolled in learning disability classes. Their

analysis comprised a minute-by-minute breakdown of the students' performance on Mature Words, Total Words Written, Words Spelled Correctly, and Correct Letter Sequences. A cumulative total for Total Words Written, Words Spelled Correctly, Mature Words, and Correct Letter Sequences was recorded for each student at the end of three, four, and five minutes. Student performance for each sample duration then was correlated with the students' Developmental Sentence Score (Lee & Canter, 1971). There was little difference in the coefficients of the three (+.80), four (+.83), and five (+.80) minute samples. Inspection of these data, however, revealed that three-minute samples provided the widest range of scores.

Additionally, the evidence from studies previously described was based on three-minute writing samples. The three-minute samples yielded data that were extremely sensitive to student change across and within grade levels, differentiated students in LD programs from students enrolled in regular classrooms.

Effectiveness Considerations

Evidence is accumulating to support the relation between academic engaged time and student achievement (cf. Graden, Thurlow, & Ysseldyke, 1982; Greenwood, Delquadri, Stanley, Terry, & Hall, 1981). It would appear, then, that the longer the measurement sample, the greater the amount of engaged time, which potentially would result in improved student gains.

Logistical Considerations

Given that formative evaluation measures are administered frequently, it seems apparent that a three-minute writing sample is

more practical than a five-minute sample. For example, if a teacher serves 15 LD children individually and measures their writing performance weekly, approximately 45 hours in the school year would be devoted to the production of five-minute writing samples. Employing three-minute writing samples reduces this time by 40% to 27 hours, without minimizing the validity of the measurement procedure.

In sum, effectiveness considerations support the use of long measurement sample, whereas logistical concerns suggest the feasibility of short samples; technical data reveal the acceptability of three, four, and five-minute samples. On the basis of these considerations, it would appear that a three-minute sample represents an acceptable compromise recommendation.

How to Measure: The Selection of Measurement Frequency

Since there are no available studies specifically investigating the effects of frequency of administration on the measurement of written expression, only a few recommendations may be advanced. First, measurement should occur frequently, at least once per week. Second, aggregating data provides more reliable information. Thus at least two, and preferably three, writing samples should be elicited on each occasion. Third, a group administered procedure is recommended on logistical grounds. While such a procedure is most time efficient, it should have no effect on either effectiveness or technical considerations.

Summary

The discussion in Chapter V supports the following recommendations:

- The behavior measured should to be Letters in Correct Sequence, Total Words Written, or Total Words Spelled Correctly
- The preferable scoring unit is either Total Words Written or Words Spelled Correctly
- The stimulus format should be a Story Starter or Topic Sentence
- Measurement should occur at least two times per week
- Each test should last for three minutes
- The samples should be collected in a group administration

Chapter VI Outline

Graphing

- Progress Measurement: Selecting a Graphing Convention
- Performance Measurement: Selecting a Graphing Convention

Data Summarization and Interpretation

- Goal-Oriented Analysis
- Program-Oriented Analysis
- Technical Considerations
- Effectiveness Considerations
- Logistical Considerations

Summary

Chapter VI

Data Utilization

Lynn S. Fuchs

The preceding chapters have focused on the top two rows of the decision matrix for three academic areas, by detailing considerations for determining "What to Measure" and "How to Measure." Having specified what will be measured and how the measurement data will be collected, one must address the bottom row of the decision matrix by determining what data-utilization procedures can be employed in a technically adequate, instructionally effective, and logistically feasible way.

Evidence suggests that teachers who collect student performance data do not necessarily use those data to make instructional decisions (Baldwin, 1976; White, 1977). Yet, procedures for interpreting student performance data appear to be an important dimension of an effective measurement and evaluation system. In Chapter I, empirical evidence (Martin, 1980; Mirkin & Deno, 1979) and a theoretical rationale (Gagne, 1965; Howell et al., 1979; Jenkins et al., 1979; Lovitt, 1977; Scriven, 1967) supporting this view were presented. This chapter further explores data utilization. First, it briefly presents methods for graphing data and reviews the technical adequacy, instructional effectiveness, and logistical feasibility of each method. Then, the chapter describes alternate procedures for data summarization and interpretation, and again reviews technical, instructional, and logistical advantages and disadvantages of those procedures.

Graphing

For progress measurement, cumulative units of curriculum per time unit are graphed; for performance measurement, successive levels of performance on samples from the same material pool per time unit are graphed. Within each measurement format, there are few alternative procedures. Further, there is little evidence for the superiority of any one graphing convention. Given this paucity of information, all three types of considerations are grouped below and discussed by basic measurement strategy.

Progress Measurement: Selecting A Graphing Convention

The progress graph was discussed in Chapter III under the heading "Selecting a Mastery Unit." As indicated there, a critical problem in progress measurement is the lack of equal intervals to represent curriculum units along the vertical axis. Therefore, a critical rationale for employing a particular graphing convention within progress measurement is how it resolves this problem.

Deno and Mirkin (1977) advocated a graphing procedure wherein mastery units along the vertical axis are plotted so that the units are spaced in accordance with the mastery time demonstrated by average students. This procedure increases the likelihood that the intervals will be equal. The graph is organized so that for the average student the level of progress is one-to-one: For each time unit, the average student is expected to master the number of pages or stories designated for that period. Average rate of progress through the curriculum, then, is depicted by a diagonal line from the lower left corner to the upper right corner of the graph (see Figure 4).

Insert Figure 4 about here

Therefore, it appears that the technical adequacy of progress measurement might be improved if the system were conceptualized as progress through pages read or words spelled of a curriculum, with the number of pages or words spaced along the ordinate axis according to the time of mastery expected of average students in the curriculum.

Logistically and instructionally, it may be advantageous to have students graph their own data. Once students are competent graphers, this procedure should reduce teacher time and therefore improve feasibility of frequent measurement. Additionally, Frumess (1973) demonstrated that students who scored and graphed their daily reading performance achieved significantly better than students who only scored their performance.

Performance Measurement: Selecting A Graphing Convention

Within performance measurement, the relative merits of equal interval and semi-logarithmic paper have been explored. In two cross-over studies, Brandstetter and Merz (1978) compared the reinforcement value of semilog graphs and linear graphs with the reinforcement value of raw scores. In the first study, reading gains made while charting on linear graphs were significantly greater than gains made while recording raw scores. In the second study, the difference between charting on semilog graphs and recording raw scores was not significant. However, because the children employed in both studies were neither randomly assigned nor even similar to each other, it is

impossible to make valid comparisons between the effectiveness of the two types of graphs.

Marston (1982) compared the prediction capabilities of both types of charts. After calculating the slope of performance for each of 82 elementary students who were measured weekly over seven weeks, predictions of student scores for weeks 8, 9, and 10 were determined. Actual student performance for weeks 8, 9, and 10 then was compared to the predictions made with equal interval and semi-logarithmic charts. For the academic areas of reading, spelling, and written expression, predictions were significantly better on the equal interval graphs.

Data Summarization and Interpretation

Once student performance data have been collected and graphed, the educator must summarize and interpret these data to determine whether the instructional program appears effective or whether that program should be changed. Two approaches to data summarization and interpretation are goal-oriented and program-oriented analyses.

Goal-Oriented Analysis

In goal-oriented data analysis, the objective is to ensure that a student's performance reaches a prespecified goal by a certain date. This goal may represent any reasonable performance level selected by the teacher. Or, in a more systematic fashion and in consonance with the principles of normalization (Wolfensberger, 1972), this goal may be a performance level commensurate with a student's mainstream peers or a level that represents a reduced discrepancy between the student's current performance and his/her age-grade appropriate level. This goal, designated the static aim (Liberty, 1972, 1975), is marked on

the graph with an X at the intersection of the desired performance level and the anticipated attainment date. Then, a line of desired progress, the dynamic aim, that connects the student's baseline median score with the static aim is drawn onto the graph.

Throughout the delivery of instrucution, data summarization consists primarily of calculating median performance within intervention periods. Data interpretation consists of the application of some form of the following rule: If on N consecutive days (i.e., 2, 3, or more) student performance data are below the dynamic aimline, then the program is judged ineffective and should be changed. Two possible consequences are: (a) a new aimline is drawn on the graph, parallel to the old aimline but originating from the intersection of the middle day on which performance was inadequate and the median performance level of those inadequate data points, and (b) a change is introduced into the student's program. This change in the program is designated on the graph with a vertical line running through the date on which the program change was introduced.

Program-Oriented Analysis

In program-oriented data analysis, the student performance level and attainment date may be specified, but are not essential to data utilization. Instead, the directive is to test changes in a student's program frequently and systematically, to move the student's performance toward the highest possible rate of improvement. One assumes that only by implementing an unending series of program changes and by comparing the effects of those program changes on a student's performance can an effective individual program emerge (Deno

& Mirkin, 1977).

Therefore, program changes are introduced regularly and are treated as experimental hypotheses to be tested by observing their effect on a student's performance. The methods of time-series analysis (Sidman, 1960) are employed to summarize and interpret student performance data. These methods are described briefly here.

Data summarization methods. Within an intervention period, four indices of student performance typically are employed to summarize data: (a) the median, a measure of central tendency representing the score that falls at the 50th percentile; (b) the split-median trend (White, 1971) or a line of fit through the data points that indicates how fast and in what direction student performance is changing; (b) the step up or down on the first day of intervention (i.e., the size and direction of the difference between the last data point of the previous intervention and the first data point of the current program); and (d) the total bounce (Pennypacker et al., 1972), which indicates the variability of the data points around the trend line.

Data interpretation methods. In time-series analysis, data interpretation is relative; one judges the effectivness of a program or experimental treatment by comparing performance among treatments. The indices of performance described above are compared across treatments to determine whether a new program has affected student performance. Therefore, a change in median, level, trend, or variability between adjacent phases and/or combinations of changes in those indices are inspected and interpreted to formulate decisions about the effectiveness of programs.

Technical Considerations

In exploring the technical strengths of goal-oriented and program-oriented data summarization and interpretation, two relevant considerations are (a) accuracy of judgments and (b) interjudge agreement. With respect to the accuracy of judgments, it appears that goal-oriented analysis is stronger (Tindal, Wesson, Mirkin, Deno, & Fuchs, 1982). Ten teachers in a rural special education cooperative were assigned randomly and then trained to use either a goal-oriented or a program-oriented analysis procedure to analyze their students' graphs. Midway through the study each teacher crossed over to the other data analysis condition. Results indicated that by the end of the study, teachers summarized data more accurately with the goal-oriented analysis rules (47% vs 12% correct summarizations), and the timing of changes in students' programs was more accurate with the goal-based rules (70% vs. 33% correctly timed changes). With respect to interjudge reliability, program-oriented analysis may be stronger. In the same study (Tindal et al., 1982), teachers' judgments with the program-oriented analysis rules were more reliable for both when program changes should be introduced (76% vs 62% agreements) and when program changes were producing student growth (88% vs 74% agreements).

The technical superiority of one data summarization/utilization method, therefore, has not been established clearly. Program-oriented analysis appears to be more reliable; goal-oriented analysis appears to be more accurate. Certainly, the differences in the results were larger and more dramatic for the reliability contrasts; goal-oriented accuracy was an average three times greater. On the basis of these

results, one might conclude tentatively that each data-utilization method has some technical strength; perhaps the goal-oriented method renders more correct, and therefore technically superior, analyses.

Effectiveness Considerations

Scant evidence exists for the superiority of either data utilization procedure in producing greater student gains. Available studies have contrasted the relative effectiveness of monitoring short-term objectives using weekly aimlines with monitoring long-term objectives using program-oriented methods (Mirkin, Fuchs, Tindal, Christenson, & Deno, 1981; Tindal, Fuchs, Christenson, Mirkin, & Deno, 1981). Results indicated that teachers believed they were more effective in the short-term objective conditions, even though there actually were no student performance differences. Perhaps the only piece of evidence directly contrasting goal-oriented and program-oriented methods supports the instructional effectivness of a goal-oriented analysis because of its effect on teacher behavior. Tindal et al. (1982) demonstrated that teachers more accurately judged effective interventions when they applied goal-oriented analysis procedures (100% vs 80% accurate judgments).

Logistical Considerations

Goal-oriented analysis also appears stronger for two logistical reasons. First, data summarization is less time consuming; it entails the computation of one rather than four statistics. Second, over two training sessions in the study described above (Tindal et al., 1982), teachers were more accurate in the goal-oriented analysis group (79% vs 68% correct decisions). Therefore, it appears that goal-oriented

analysis methods are more feasible because they are less time consuming during both training and day-to-day implementation.

Nevertheless, evidence suggests that teachers may prefer a combination of the two data-utilization methods. Fuchs, Wesson, Tindal, Mirkin, and Deno (1982) found that teachers preferred the goal-oriented approach for (a) monitoring progress toward IEP goals, (b) the ease of its use, (c) its efficiency, (d) a guide when to change a student's instructional program, (e) the ease with which it could be described to parents and teachers, (f) its more adequate representation of student performance, and (g) its overall usefulness. The program-oriented approach was preferred by most teachers only as a guide for what to change in a student's instructional program. When asked to name the data-utilization system of their choice, one-half of the surveyed teachers indicated that they preferred to use a combination of the two approaches. Therefore, despite the teachers' overwhelming preference for goal-oriented evaluation, many chose a combination of the two approaches. This finding may be attributed to the fact that goal-oriented evaluation addresses the question of when, not what, to change in a student's program, and that teachers preferred program-oriented evaluation for determining what to change in an educational plan. For handicapped children, the question "what to change" may be especially problematic, and this may have led some teachers to conclude that a combination of the two strategies is optimal.

A strong experimental contrast of the two data-utilization strategies, one with dramatic and persuasive results, currently is not

available. Nevertheless, evidence presented in this chapter suggests that a goal-oriented analysis may be more technically adequate, more feasible, more efficient, and more instructionally useful. Given these results, along with teachers' preference for a combination of the two data-utilization approaches, a combined data-utilization method that borrows more heavily from the goal-oriented approach is recommended. In this approach (Mirkin, Deno, Fuchs, Wesson, Tindal, Marston, & Kuehnle, 1981), teachers draw the dynamic aimline on the graph. Then, a split-median trend line on 7 to 10 student performance data points is graphed and compared to the slope of the dynamic aimline. If the student performance slope is less steep than the aimline, a program change is introduced. Logistically, this data-utilization rule is facilitated by the availability of a soft-ware computer package whereby teachers can enter student performance data and access a student graph with aimline and a decision concerning whether a program change is indicated. Future research should investigate the relative adequacy and instructional effectiveness of the combined evaluation strategy.

Summary

The discussion in Chapter IV supports the following recommendations:

- Within progress measurement, a small unit of mastery, such as pages in a book mastered or words in a curriculum spelled, should be graphed.
- Within performance measurement, equal interval graphing paper should be used.
- A combination of goal- and program-oriented data summarization and analyses should be employed whereby a split-median trend is drawn through 7 to 10 data points,

and if that trend is flatter than the goal line, a program modification is introduced.

Chapter VII Outline

What to Measure: The Selection of a Behavior

How to Measure: The Selection of a Basic Strategy

How to Measure: The Selection of a Score, a Difficulty Level, and a Measurement Domain

How to Measure: The Selection of a Measurement Frequency and a Sample Duration

How to Measure: The Selection of a Criterion of Mastery or Goal

How to Measure: The Selection of a Procedure for Generating Test Samples

How to Measure: The Selection of Administration and Scoring Procedures

Goal and Objective Form
Measurement System Form

How to Use Data

Chapter VII

A Case Study

Lynn S. Fuchs

After presenting a rationale for direct and continuous evaluation in Chapter I, and research supporting its use, a framework for developing useful, adequate, and feasible measurement and evaluation procedures was developed in Chapter II. In Chapters III through VI, that decision framework was employed. Within that decision matrix, available research was reviewed; from that review, procedures were recommended for measuring students' performance in reading, writing, and spelling, and for using those measurement data to formulate decisions about the adequacy of pupil progress and programs.

Concurrent with this effort to integrate available research into a meaningful framework in order to specify how one might best measure and evaluate student performance, a manual (Mirkin, Deno, Fuchs, Wesson, Tindal, Marston, & Kuehnle, 1981) was developed to train teachers to implement the procedures recommended in this monograph. This chapter presents a case study of one teacher's implementation of the procedures described in the manual. This case study is presented as the concluding chapter of this monograph to provide the reader with a concrete notion of how one might create and employ a measurement and evaluation system.

This case study describes how Mrs. R. measured and evaluated the reading progress of Michael, a mildly handicapped fourth grader who was reading at a second grade level. The case study is structured to match the organization of the decision matrix developed in this

monograph.

What to Measure: The Selection of a Behavior

In compliance with the procedures recommended in this monograph, Mrs. R. decided that she would measure Michael's reading aloud from text. Reading aloud from text demonstrates acceptable slope as well as construct and criterion validity. Measuring reading aloud in context provides rich information for making sound decisions with which programs can be enhanced; such measurement is feasible relative to the measurement of other reading behaviors.

How to Measure: The Selection of a Basic Strategy

Mrs. R. decided that her basic measurement strategy would be performance measurement. In other words, she decided to select one level of difficulty for the reading selections on which she would measure Michael's progress; her goal was to improve Michael's performance on that material. Since there was no clear advantage to either progress or performance measurement, Mrs. R. chose performance measurement because of her personal preference.

How to Measure: The Selection of a Score, a Difficulty Level, and a Measurement Domain

Mrs. R. chose correct and error rate as the scores she would monitor. She selected Level 2 of the SRA Series as the material from which she would draw selections for measuring Michael's reading aloud in context behavior; Level 2 was chosen because it represented a mid-range difficulty for Michael. (He initially read 30 words per minute correct with no more than 11 errors.) She decided to monitor correct and error rate because the correct rate should represent technically

adequate data; the error rate should supplement the correct rate by providing accuracy information.

Having decided on correct and error rate as well as on Level 2 of the SRA series, Mrs. R. was ready now to determine what the size of the measurement domain would be. She decided on a mid-sized domain, all of Level 2, because it was likely to render data with relatively low variability and with an acceptable slope; further, it would probably remain an appropriate difficulty level for Michael over the entire school year.

How to Measure: The Selection of a Measurement Frequency and a Sample Duration

Given the technical, logistical, and effectiveness considerations discussed in the monograph, the following measurement procedures were recommended in the manual: a schedule of at least twice weekly and a sample duration of one minute. Mrs. R. decided to adhere to these recommendations.

How to Measure: The Selection of a Criterion of Mastery or Goal

Given the discussion in the monograph and the procedural recommendations described in the manual, Mrs. R. decided on a mastery criterion of 80 words per minute with no more than 8 errors (or a 90% accuracy criterion). This criterion of mastery or goal represented to Mrs. R. a reasonable but ambitious amount of improvement for Michael. Additionally, it fell close to the recommended rates.

How to Measure: The Selection of a Procedure for Generating Test Samples

Mrs. R. decided on random selection for generating test samples because of the theoretical advantages of such a selection procedure discussed in the monograph. To implement a random procedure for selecting passages, she followed the directions presented in the manual: (a) Use passages selected from the level that represents the annual goal, and write on equal size slips of paper the number of each of the pages in those stories that do not have excessive dialogue, indentations, and/or unusual pronouns; (b) Put the slips of paper into a drawbag and shake it; (c) Randomly pick a slip of paper; and (d) Have the student begin reading on the page number shown on the slip of paper.

How to Measure: The Selection of Administration and Scoring Procedures

Once the practitioner has defined a measurement system thus far, there are only a few alternatives for administering and scoring tests. The two primary considerations in choosing among alternatives are: (a) technical, that is, maintaining consistent procedures across testing occasions, and (b) logistical, that is, designing efficient administration and scoring procedures. Mrs. R. decided to adhere to the following procedures described in the manual: (a) Put the student copy in front of and facing the student; (b) Say to the student: "When I say 'start,' begin reading aloud at the top of this page. Try to read each word. If you wait for a word too long, I'll tell you the word. You can skip words that you don't know. At the end of one

minute, I'll say, 'stop.'" (Give student 3 seconds before supplying list); (c) Turn on the stopwatch as you say "start"; (d) Follow along on another copy circling with a pencil incorrectly read words; (e) At one minute, say "stop" and turn off the stopwatch; (f) Place a slash after the last word read; (g) Count the number of words correct and the number of errors.

Having specified "What to Measure" and "How to Measure," Mrs. R. was ready to complete the following Goal and Objective Form and the following Measurement System Form.

Goal and Objective Form

GOAL In 19 weeks (# school weeks until year's end), when presented with stories from Level 2-SRA Series (Level #, series), Michael (student's name), will read aloud at the rate of 80 (words per minute correct) with no more than 8 (#) errors.

OBJECTIVE Each successive week, when presented with a random selection from Level 2 - SRA Series (same level # and series as above), the student will read aloud at an average increase of 2.6 (#) words per minute and no increase in errors.

Measurement System Form

BEHAVIOR: reading aloud in context

FREQUENCY: at least twice weekly

DURATION OF TEST: one minute

DIFFICULTY LEVEL: Level 2, SRA Series

SIZE OF DOMAIN: all of Level 2

TEST ADMINISTRATION PROCEDURE: see Manual

SCORING PROCEDURE: see Manual

How to Use Data

Having specified what she would measure and how she would measure, Mrs. R. had to determine how she would use the data she collected. The first decision Mrs. R. made was to graph the data on equal interval paper. Figure 4 displays Michael's graph with "Words Read Per Minute" on the vertical axis and "School Days" on the horizontal axis. The first three data points on this figure indicate Michael's baseline performance on the Level 2 material. The vertical lines following these baseline data indicate the introduction of new dimensions into Michael's reading program. These program dimensions are labeled briefly at the top of these lines and described in more detail on the Instructional Change Form (see Figure 5). The large X on the graph indicates the mastery criterion or goal that Mrs. R. set for Michael; the diagonal line from the baseline median to the X is Michael's dynamic aimline, which depicts the rate of progress Michael had to exhibit in order to meet his goal as anticipated.

Insert Figure 5 about here

With this graph established, Mrs. R. could plot data points and easily see, on any given day, how Michael's performance compared to his dynamic aim, or the level at which Michael had to perform in order to reach the long-term goal. Mrs. R., then, decided to adopt the

data-utilization rule established in the monograph and recommended in the manual: If a split-median trend line drawn through 7 to 10 data points is greater than or equal to the slope of the dynamic aimline, maintain the student's program; if the trend line is less than the aimline slope, introduce a change into the student's program. As Figure 4 illustrates, Michael's performance improved dramatically over his previous performance with the introduction of the third program change.

As Mrs. R. used the monograph decision framework and manual procedures to formulate this system, she established a close connection between the instruction she provided Michael and the way she measured and evaluated his progress. With such a measurement and evaluation system, Michael's educational program and progress toward goals was evaluated formatively. In response to measurement data, Michael's program was modified throughout the treatment phase to improve the likelihood that Michael would achieve his annual goal. In a similar way, practitioners can employ the decision matrix and the integrative discussion presented in this monograph to create a useful, feasible, and adequate measurement and evaluation system.

References

Alper, T. G., & White, O. R. Precision teaching: A tool for the school psychologist and teacher. Journal of School Psychology, 1971, 9(4), 445-454.

Artley, A. S. A study of certain relationships existing between general comprehension and reading comprehension in a specific subject matter area. Journal of Educational Research, 1944, 37, 364-373.

Askov, E., Otto, W., & Fischbach, T. Development of specific reading skills in adult education. In F. P. Greene (Ed.), Reading: The right to participate. Milwaukee: National Reading Conference, 1971.

Ayllon, T., Garber, S., & Pisor, K. Reducing time limits: A mean to increase behavior of retardates. Journal of Applied Behavior Analyses, 1976, 9, 247-252.

Baldwin, V. Curriculum concerns. In M. A. Thomas (Ed.), Hey, don't forget about me. Reston, VA: Council for Exceptional Children, 1976.

Beldin, H. L. Informal reading testing: Historical review and review of the research. In W. Durr (Ed.), Reading difficulties: Diagnosis, correction, and remediation. Newark, Del: International Reading Association, 1970.

Biemiller, A. The development of the use of graphic and contextual information as children learn to read. Reading Research Quarterly, 1970, 6(1), 75-96.

Bohannon, R. Direct and daily measurement procedures in the identification and treatment of reading behaviors of children in special education. Unpublished doctoral dissertation, University of Washington, 1975.

Botel, M. A comparative study of the validity of the Botel Reading Inventory and selected standardized tests. International Reading Association, Conference Proceedings, Part 1, 1968, 13, 722, 727.

Bowers, J. E. End-of-year evaluation report for the 1979-80 Title I Neglected and Delinquent Program of the Minneapolis Public Schools. Unpublished manuscript, 1980. (Available from Educational Services Division, Minneapolis Public Schools, 807 N.E. Broadway, Minneapolis, MN 55413).

Bradley, J. M. Evaluating reading achievement for placement in special education. Journal of Special Education, 1977, 10(2), 168-177.

Brandstetter, G., & Merz, C. Charting scores in precision teaching for skill acquisition. Exceptional Children, 1978, 45(1), 42-48.

Cartwright, C. P. Written expression and spelling. In R. M. Smith (Ed.), Teacher diagnoses of educational difficulties. Columbus, OH: Charles E. Merrill, 1969.

Churchman, D., Petrosko, J., Spooner-Smith, L. The theoretical basis for formative evaluation. Paper presented at the annual meeting of the American Education Research Association, Washington, D.C., April 1975. (ERIC Document Reproduction Service No. ED 105 567)

Cohen, M. Teacher judgment concerning arithmetic disabilities accounted for by a functional assessment battery. Unpublished doctoral dissertation, University of Washington, 1975.

Conant, M. M. The construction of a diagnostic reading test. New York: Teachers College, Columbia University. Contributions to Education No. 861, 1942.

Davis, F. B. Fundamental factors of comprehension in reading. Psychometrika, 1944, 9, 185-197.

Davis, F. B. A brief comment on Thurstone's note on a reanalysis of Davis' reading tests. Psychometrika, 1946, 11, 249-255.

Deno, S. L., Chiang, B., Tindal, G., & Blackburn, M. Experimental analysis of program components: An approach to research in CSDC's (Research Report No. 12). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1979. (ERIC Document Reproduction Service No. ED 185 756)

Deno, S. L., Marston, D., & Mirkin, P. K. Valid measurement procedures for continuous evaluation of written expression. Exceptional Children, 1982, 48, 368-371.

Deno, S., Marston, D., Mirkin, P., Lowry, L., Sindelar, P., & Jenkins, J. The use of standard tasks to measure achievement in reading, spelling, and written expression: A normative and developmental study (Research Report No. 87). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1982.

Deno, S. L., & Mirkin, P. K. Data based program modification: A manual. Arlington, VA: The Council for Exceptional Children, 1977.

Deno, S. L., & Mirkin, P. K. Data-based IEP development: An approach to substantive compliance. Teaching Exceptional Children, Spring 1980, 4-9.

Deno, S. L., Mirkin, P. K., Chiang, B., & Lowry, L. Relationships among simple measures of reading and performance on standardized achievement tests (Research Report No. 20). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1980. (ERIC Document Reproduction Service No. ED 197 507)

Deno, S. L., Mirkin, P. K., Lowry, L., & Kuehnle, K. Relationships among simple measures of spelling and performance on standardized achievement tests (Research Report No. 21). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1980. (ERIC Document Reproduction Service No. ED 197-508)

Deno, S. L., Mirkin, P. K., & Marston, D. Relationships among simple measures of written expression and performance on standardized achievement tests (Research Report No. 22). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1980. (ERIC Document Reproduction Service No. ED 197 509)

Dunn, L. M., & Markwardt, F. C. Peabody individual achievement test. Circle Pines, MN: American Guidance Service, 1970.

Dunn, L. M., & Smith, J. O. Peabody language development kit. Circle Pines, MN: American Guidance Service, 1967.

Epstein, S. The stability of behavior, II: Implications for psychological research. American Psychologist, 1980, 35(9), 790-806.

Eurick, A. C. The relationship of speed of reading to comprehension. School and Society, 1930, 32, 404-406.

Freyberg, P. S. The concurrent validity of two types of spelling test. British Journal of Educational Psychology, 1970, 40, 68-71.

Frumess, S. A comparison of management groups involving the use of the standard behavior chart. Unpublished doctoral dissertatio University of Texas, 1973.

Fuchs, L. S. The concurrent validity of progress measures of basal reading material. Unpublished doctoral dissertation, University of Minnesota, 1981.

Fuchs, L., & Deno, S. The relationship between curriculum-based mastery measures and standardized achievement tests in reading (Research Report No. 57). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1981. (ERIC Document Reproduction Service No. ED 212 662)

Fuchs, L. S., Deno, S. L., & Marston, D. Use of aggregation to improve the reliability of simple direct measures of academic performance (Research Report No. 93). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1982.

Fuchs, L. S., Deno, S. L., & Roettger, A. The effect of measurement strategy on spelling achievement: An N of 1 study. Unpublished manuscript, 1980. (Available from L. Fuchs, 46 Chamberlain Parkway, Worcester, MA 01602)

Fuchs, L. S., Tindal, G., & Deno, S. L. Effects of varying item domain and sample duration on technical characteristics of daily measures in reading (Research Report No. 48). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1981. (ERIC Document Reproduction Service No. ED 211 606)

Fuchs, L., Wesson, C., Tindal, G., Mirkin, P., & Deno, S. Teacher efficiency in continuous evaluation of IEP goals (Research Report No. 53). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1981.

Fuchs, L., Wesson, C., Tindal, G., Mirkin, P., & Deno, S. Instructional changes, student performance, and teacher preferences: The effects of specific measurement and evaluation procedures (Research Report No. 64). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1982.

Gagne, R. The conditions of learning. New York: Holt, Reinhart, & Winston, 1965.

Gates, A. I. An experimental and statistical study of reading and reading tests. Journal of Educational Psychology, 1921, 12(6), 303-313, 378-391, 445-464.

Graden, J., Thurlow, M., & Ysseldyke, J. Academic engaged time and its relationship to learning: A review of the literature (Monograph No. 17). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1982. (ERIC Document Reproduction Service No. ED 214 930)

Gray, W. S. Summary of investigations relating to reading. Supplementary Educational Monograph No. 28, 1925.

Greenwood, C., Delquadri, J., Stanley, S., Terry, B., & Hall, R. Process-product study of relationships among instructional ecology, student response, and academic achievement. Unpublished manuscript, Juniper Gardens Children's Project, University of Kansas, 1981.

Guszak, F. J. A comparative study of the validity of the cloze test and the Metropolitan Achievement Test (Reading Comprehension Subtest) for making judgments of instructional levels. Unpublished manuscript, 1969, Austin: Texas University.

Hall, W. E., & Robinson, F. P. An analytical approach to study of reading skills. Journal of Educational Psychology, 1945, 36, 429-442.

Hammill, D. D., & Larsen, S. C. The test of written language. Austin, TX: PRO-ED, 1978.

Hammill, D. D., & Noone, J. Improving spelling skills. In D. Hammill and N. Bartel (Eds.), Teaching children with learning and behavior problems. Boston: Allyn & Bacon, 1975.

Haring, N. G., & Gentry, N. D. Direct and individualized instructional procedures. In N. G. Haring & R. Schiefelbusch (Eds.), Teaching special children. New York: McGraw-Hill, 1976.

Haring, N., & Krug, D. A. Placement in regular programs: Procedures and results. Exceptional Children, 1975, 41(6), 413-417.

Haring, N., Liberty, K., & White, O. Instructional hierarchies research project: Handbook of experimental procedures. Unpublished manuscript, Seattle: University of Washington, undated.

Haring, N. G., & Lovitt, T. C. The application of functional analysis of behavior by teachers in a natural school setting (Final Report). Seattle: University of Washington, Experimental Education Unit, 1969. (ERIC Reproduction Service No. ED 045 598)

Haring, N., Maddux, L. & Krug, D. A. Investigation of systematic instructional procedures to facilitate academic achievement in mentally retarded disadvantaged children (Final Report). Seattle: University of Washington, Experimental Education Unit, 1972.

Haring, N. G., White, O. R., & Liberty, K. A. Field-initiated research studies: An investigation of learning and instructional hierarchies in severely and profoundly handicapped children (Annual Report, 1978-1979). Seattle: University of Washington, Child Development and Mental Retardation Center, Experimental Education Unit, 1979.

Harris, A. J. & Jacobson, M. D. Basic elementary vocabularies - The first R series. New York: MacMillan, 1972.

Haughton, E. Aims: Growing and sharing. In J. Jordan & L. Robbins (Eds.), Let's try doing something else kind of thing. Arlington, VA: The Council for Exceptional Children, 1972.

Hays, W. Statistics of the social sciences (2nd ed.). New York: Holt, Rinehart, & Winston, 1973.

Hersen, M., & Barlow, D. H. Single case experimental designs: Strategies for studying change. New York: Pergammon Press, 1976.

Hildreth, G. Teaching spelling. New York: Henry Holt, 1955.

Horn, E. Spelling. In W. S. Monroe (Ed.), Encyclopedia of educational research. New York: MacMillan, 1941.

Horn, E. Research in spelling. Elementary English Review, 1944, 21, 6-13.

Horn, E. Phonics and spelling. Journal of Education, 1954, 136, 233-235, 246.

Howell, K., Kaplan, J. R. & O'Connell, C. Y. Evaluating exceptional children: A task analytic approach. Columbus, OH: Charles E. Merrill, 1979.

Hunt, K. W. Recent measures in syntactic development. Elementary English, 1966, 42, 732-739.

Jenkins, J. R., Deno, S. L., & Mirkin, P. K. Measuring pupil progress toward the least restrictive environment (Monograph No. 10). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1979. (ERIC Document Reproduction Service No. ED 185 767)

Judd, C. H. Measuring the work of the public schools. Cleveland Educational Survey, 1916, 124-161.

Karlsen, B., Madden, R. R., Gardner, E. F. Stanford diagnostic reading test (Green Level Form B). New York: Harcourt Brace Jovanovich, 1975.

Kelley, T. L. Interpretation of educational measurements. Yonkers-on-Hudson, NY: World Book, 1927.

Larsen, S. C., & Hammill, D. D. Test of written spelling. Austin: Empiric Press, 1976.

Lee, L., & Canter, S. M. Developmental sentence scoring. Journal of Speech and Hearing Disorders, 1971, 36, 335-340.

Liberty, K. A. Data decision rules (Working Paper No. 20). Eugene, Oregon: University of Oregon, Regional Resources Center, 1972.

Liberty, K. A. Decide for progress: Dynamic aims and data decisions. Seattle: University of Washington, Experimental Education Unit, Child Development and Mental Retardation Center, 1975.

Lindsley, O. R. Direct measurement and prosthesis of retarded behavior. Journal of Education, 1964, 147, 62-81.

Lindsley, O. R. Precision teaching in perspective: An interview with Ogden R. Lindsley. Teaching Exceptional Children, 1971, 3(3), 114-119.

Lovitt, T. In spite of my resistance, I've learned from children. Columbus, OH: Charles E. Merrill, 1977.

Madden, R., Gardner, E. F., Rudman, H. C., Karlsen, B., & Merwin, J. C. Stanford achievement test. New York: Harcourt Brace Jovanovich, 1978.

Marston, D. B. The technical adequacy of direct, repeated measurement of academic skills in low achieving elementary students. Unpublished doctoral dissertation, University of Minnesota, 1982.

Marston, D., & Deno, S. The reliability of simple, direct measures of written expression (Research Report No. 50). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1981. (ERIC Document Reproduction Service No. ED 212 663)

Martin, M. D. A comparison of variations in data utilization procedures on the reading performance of mildly handicapped students. Unpublished doctoral dissertation, University of Washington, 1980.

McCracken, R. A. Standardized reading tests and informal reading inventions. Education, 1962, 82, 366-369.

Messick, S. Test validity and the ethics of assessment. American Psychologist, 1980, 35(11), 1012-1027.

Mirkin, P. A comparison of the effects of three evaluation strategies and contingent consequences on reading performance. Unpublished doctoral dissertation, University of Minnesota, 1978.

Mirkin, P. K., & Deno, S. L. Formative evaluation in the classroom: An approach to improving instruction (Research Report No. 10). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1979. (ERIC Document Reproduction Service No. ED 185 754)

Mirkin, P., Deno, S., Fuchs, L., Wesson, C., Tindal, G., Marston, D., & Kuehnle, K. Procedures to develop and monitor procedures toward progress on IEP goals. Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1981.

Mirkin, P., Deno, S., Tindal, G., & Kuehnle, K. Formative evaluation: Continued development of data utilization systems (Research Report No. 23). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1980. (ERIC Document Reproduction Service No. ED 197 510)

Mirkin, P., Fuchs, L., Tindal, G., Christenson, S., & Deno, S. The effect of IEP monitoring strategies on teacher behavior (Research Report No. 62). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1981.

Morrissey, P. A., & Semmel, M. I. Instructional models for the learning disabled. Theory into Practice, 1976, 14, 110-122.

Myklebust, H. R. Development and disorders of written language. New York: Grune & Stratton, 1965.

Nunnally, J. C. Tests and measurements: Assessment and prediction. New York: McGraw-Hill, 1959.

Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1978.

Oliver, J., & Arnold, R. D. Comparing a standardized test, an informal reading inventory and teacher judgment on third grade reading. Reading Improvement, 1978, 15(1), 56-59.

Patterson, D. G., & Tinker, M. A. Time limit vs. work limit methods. American Journal of Psychology, 1930, 42, 101-104.

Pennypacker, H. S., Koenig, C. H., & Lindsley, O. R. Handbook of the standard behavior chart (prelim. ed.). Kansas City, KS: Precision Media, 1972.

Powell, W. K. Validity of the IRI reading levels. Elementary English, 1971, 48, 637-642.

Pressey, L. W., & Pressey, S. L. A critical study of the concept of silent reading ability. Journal of Educational Psychology, 1921, 12, 35-39.

Quilling, M., & Otto, W. Evaluation of an objective based curriculum in reading. Journal of Educational Research, 1971, 65(1), 15-18.

Rowell, C. G. Don't throw away those spelling test papers... yet! Elementary English, 1975, 52(2), 253-257.

Samuels, S. J. Some essentials of decoding. Exceptional Education Quarterly, 1981, 2(1), 11-26.

Sassenrath, J. M. Alpha factor analyses of reading measures at the elementary, secondary, and college levels. Journal of Reading Behavior, 1972, 5, 304-315.

Sidman, M. Scientific research: Evaluating experimental data in psychology. New York: Basic Books, 1960.

Scriven, M. The methodology of evaluation. In R. Tyler, R. Gagne, R. M. Scriven (Eds.), Perspectives of curriculum evaluation. American Educational Research Association Monograph Series on Curriculum Evaluation. Chicago: Rand McNally, 1967.

Skinner, B. F. The behavior of organisms: An experimental analysis. New York: Appleton-Century-Crofts, 1938.

Stake, R. E. Objectives, priorities, and other judgment data. Review of Educational Research, 1970, 40, 181-212.

Stanley, J. C. Reliability. In R. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971.

Starlin, C. Evaluating and teaching reading to "irregular" kids. Iowa Perspective, Iowa Department of Public Instruction, December, 1979, 1-11.

Starlin, C., & Starlin, A. Guidelines for continuous decision making. Bemidji, MN: Unique Curriculums Unlimited, 1974.

Stoker, H., & Kropp, R. The predictive validities and factorial context of the Florida state-wide ninth-grade testing program battery. Florida Journal of Educational Research, 1960, 105-114.

Thurstone, L. L. Note on a reanalysis of Davis' reading tests. Psychometrika, 1946, 11, 185-188.

Tindal, G., & Deno, S. L. Daily measurement of reading: Effects of varying the size of the item pool (Research Report No. 55). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1981. (ERIC Document Reproduction Service No. ED 211 605)

Tindal, G., Fuchs, L., Christenson, S., Mirkin, P., & Deno, S. The relationship between student achievement and teacher assessment of short- or long-term goals (Research Report No. 61). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1981.

Tindal, G., Wesson, C., Mirkin, P., Deno, S., & Fuchs, L. Comparison of goal-oriented and program-oriented data utilization procedures (Research Report, in preparation). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1982.

Tinker, M. A. Photographic measures of reading ability. Journal of Educational Psychology, 1929, 20, 184-191.

Tinker, M. A. Speed vs. comprehension in reading as affected by level of difficulty. Journal of Educational Psychology, 1939, 30, 81-94.

Traxler, A. E. A study of the Van Wagenen-Dvorak Diagnostic Examination of Silent Reading Abilities. Educational Records Bulletin No. 31. New York: Educational Records Bureau, 1941, 33-41.

Van Etten, C., & Van Etten, G. The measurement of pupil progress and selecting instructional materials. Journal of Learning Disabilities, 1976, 9(8), 469-480.

Videen, J., Deno, S., & Marston, D. Correct word sequences: A valid indicator of proficiency in written expression (Research Report No. 84). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities, 1982.

Wallace, G., & Larsen, S. C. Educational assessment of learning problems: Testing for teaching. Boston: Allyn & Bacon, 1978.

White, O. R. A pragmatic approach to the description of progress in the single case. Unpublished doctoral dissertation, University of Oregon, 1971.

White, O. R. Evaluating the educational process (Working Paper). Seattle: University of Washington, Child Development and Mental Retardation Center, Experimental Education Unit, 1974.

White, O. R. Behaviorism in special education: An area for debate. In R. D. Kneedler & S. G. Tarver (Eds.), Changing perspectives in special education. Columbus, OH: Charles E. Merrill, 1977.

White, O. R., & Haring, N. G. Exceptional teaching: A multimedia training package. Columbus, OH: Charles E. Merrill, 1976.

White, O. R., & Haring, N. G. Exceptional teaching (2nd ed.). Columbus, OH: Charles E. Merrill, 1980.

Wolfensberger, W. The principle of normalization in human services. Toronto: National Institute on Mental Retardation, 1972.

Woodcock, R. W. Woodcock reading mastery tests (Form A). Circle Pines, MN: American Guidance Service, 1973.

Table 1

The Measurement Model[a]

Frequency of Measurement	Directness of Measurement: Indirect	Directness of Measurement: Direct
Infrequent	Type I	Type III
Frequent	Type II	Type IV

[a]Model was adapted from "The Measurement of Pupil Progress and Selecting Instructional Materials" by C. Van Etten and G. Van Etten, Journal of Learning Disabilities, 1976, 9(8), 469-480.

Table 2

Decision-Making Matrix[a]

	Technical	Effectiveness	Logistical
What to Measure	T-1	E-1	L-1
How to Measure	T-2	E-2	L-2
How to Use Data	T-3	E-3	L-3

[a]Numbers are given to label cells. Corresponding questions are presented in Table 3.

Table 3

Questions Posed in the Decision-Making Matrix

Considerations	
Technical	
T-1.	What to measure: What behaviors validly and reliably index growth and are sensitive to the effects of instruction in the domains of reading, spelling, and written expression?
T-2.	How to measure: What measurement procedures render reliable, valid, and sensitive representations of student achievement?
T-3.	How to use data: What data summarization methods, graphing conventions, and data interpretation procedures are statistically and/or psychometrically acceptable?
Effectiveness	
E-1	What to measure: In a given domain the measurement of which behavior results in improved student growth?
E-2	How to measure: What measurement procedures positively influence the rate of student improvement?
E-3	How to use data: What approaches to evaluation result in more successful programs?
Logistical	
L-1.	What to measure: In a given domain, what behaviors can be repeatedly and easily administered by teachers without excessive time demands?
L-2.	How to measure: What measurement procedures allow a teacher to reduce time engaged in measurement activities?
L-3.	How to use data: In which data summarization methods, graphing conventions, and data interpretation procedures can teachers be effectively and efficiently trained, and which procedures are less time consuming?

Table 4

Differences Between Progress and Performance Measurement

Type of Measure	Difficulty Level of Measurement Material	Goal
Performance	remains constant	to improve performance on same level of material.
Progress	increases through a skills sequence	to improve rate of progress through increasingly more difficult material.

Figure 1

Decision Flow for Academic Domains of Reading and Spelling

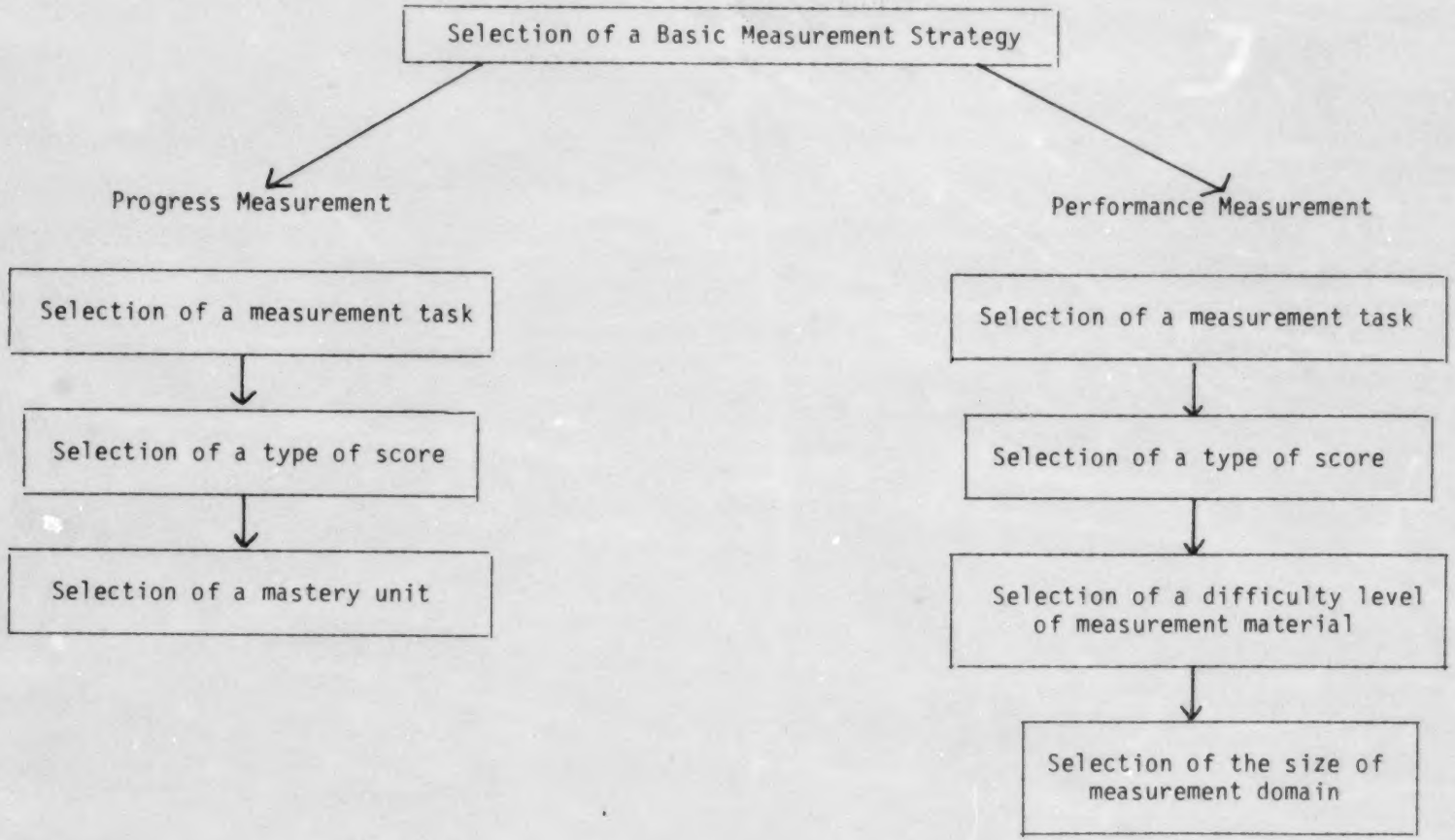
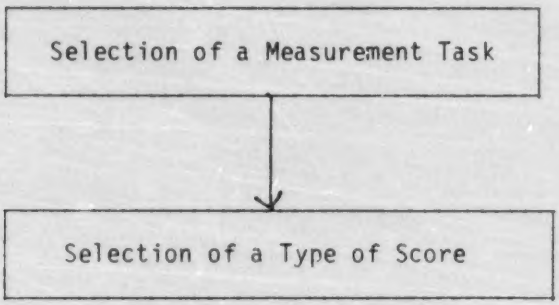

Selection of a Basic Measurement Strategy
Progress Measurement
Performance Measurement
Selection of a measurement task
Selection of a type of score
Selection of a mastery unit
Selection of a measurement task
Selection of a type of score
Selection of a difficulty level of measurement material
Selection of the size of measurement domain

Figure 2

Decision Flow for Academic Domain of Written Expression



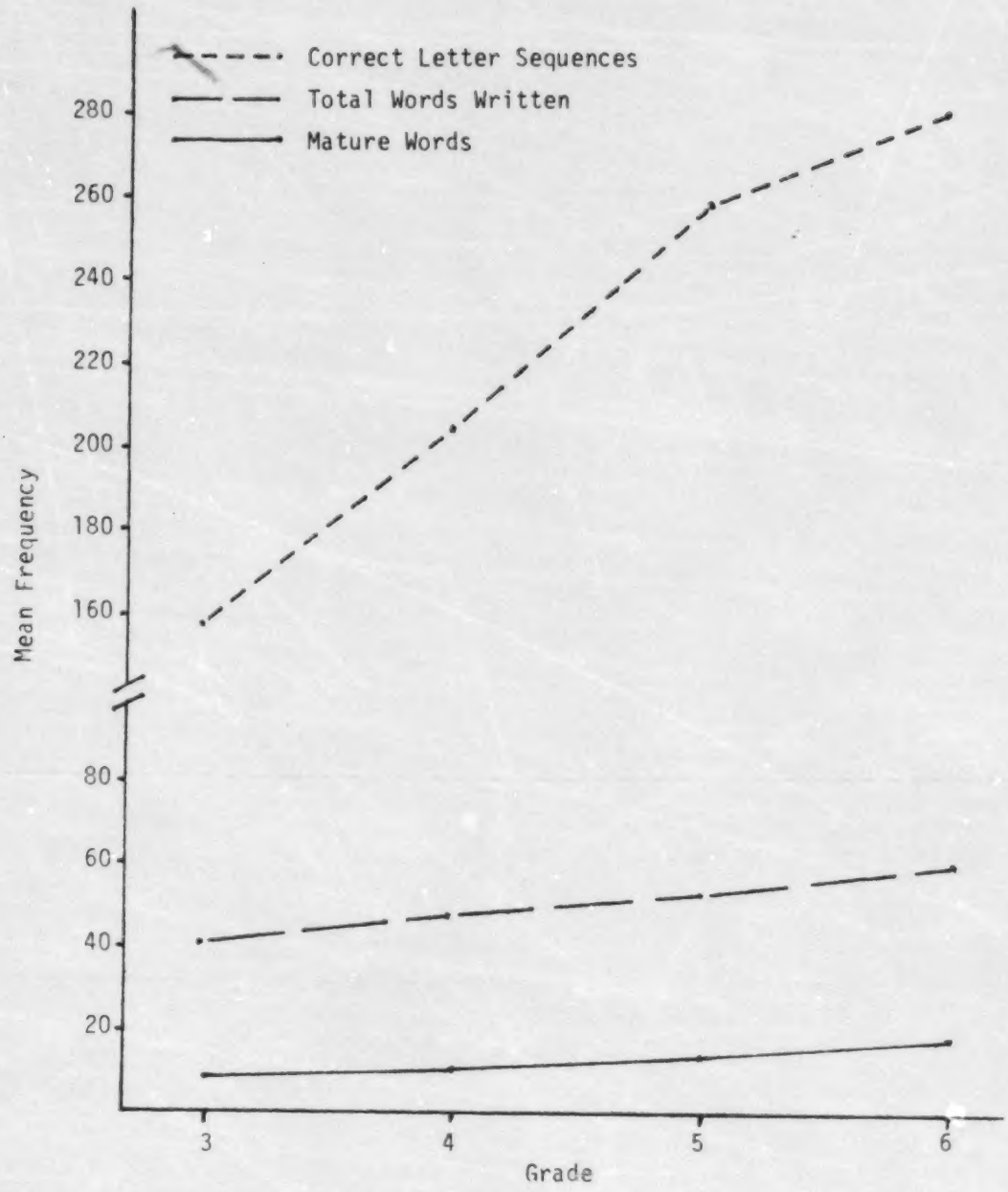


Figure 3. Mean Performance on Three Measures of Written Expression.

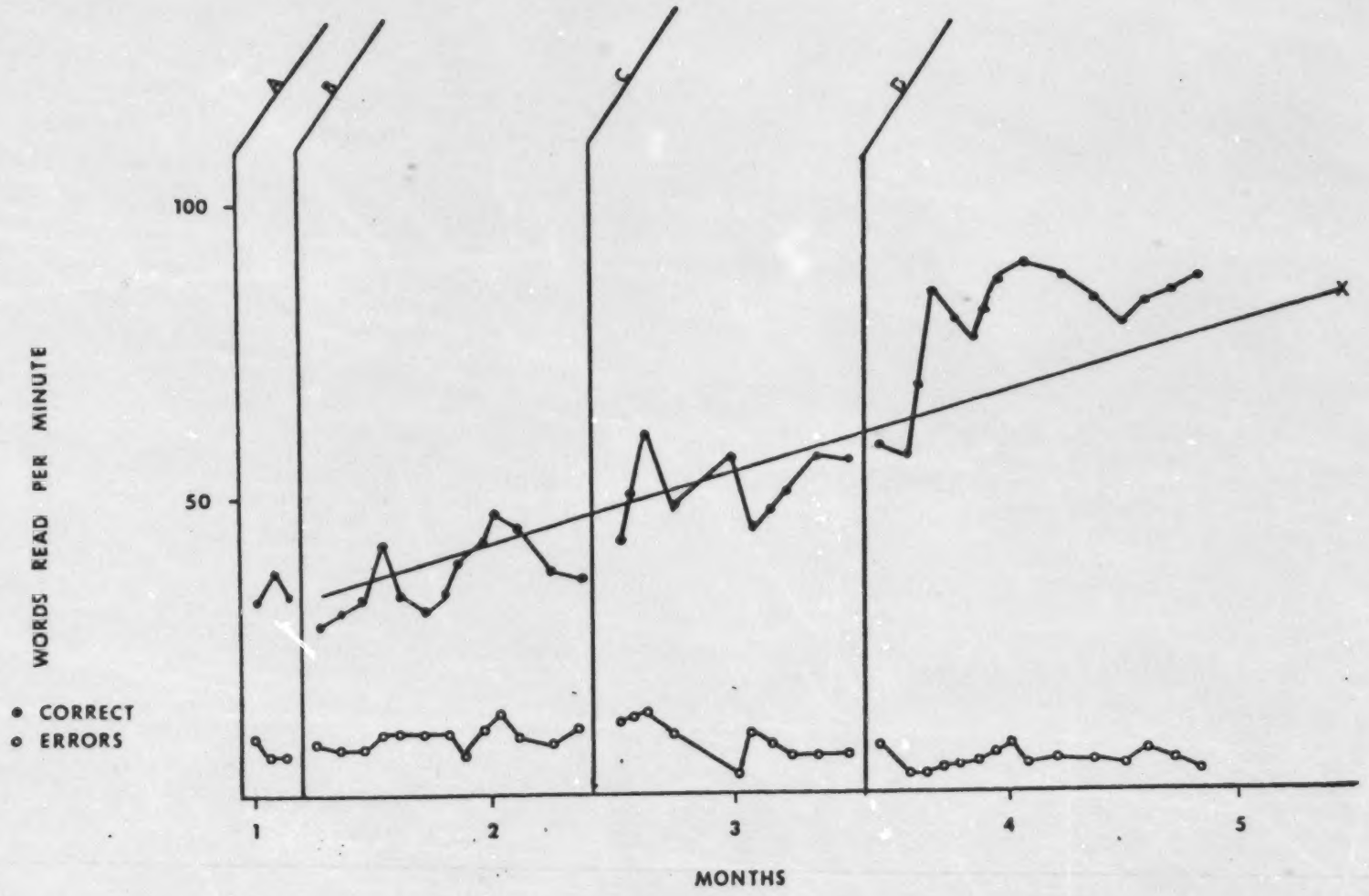


Figure 4. Number of Correct Words (•) and Errors (o) Per Minute Read by Michael from Pages in SRA, Level 2 Across Time, Under Baseline (A) and Three Instructional Strategies (B, C, and D).

Instructional Change Form

Instructional Procedures	Arrangement	Time	Materials	Motivational Strategies
Oral Reading Practice Comprehension exercises	Group (1:5)	45 minutes	Double Action Short Story, Part 2 Story Writing & class discussion	Generating own stories
Language Experience Approach	Individual with para-professional	same	Student's own stories File cards Story Folder	same
Language Experience	Individual with para-professional	20 minutes	See above	same
Reading Comprehension Activities	Individual with teacher	20 minutes	McCall-Crabbs, Book E SRA kit	Individual arrangement with teacher

Figure 5. Michael's Instructional Change Form

PUBLICATIONS

Institute for Research on Learning Disabilities
University of Minnesota

The Institute is not funded for the distribution of its publications. Publications may be obtained for $3.00 per document, a fee designed to cover printing and postage costs. Only checks and money orders payable to the University of Minnesota can be accepted. All orders must be prepaid.

Requests should be directed to: Editor, IRLD, 350 Elliott Hall; 75 East River Road, University of Minnesota, Minneapolis, MN 55455.

Ysseldyke, J. E. Assessing the learning disabled youngster: The state of the art (Research Report No. 1). November, 1977.

Ysseldyke, J. E., & Regan, R. R. Nondiscriminatory assessment and decision making (Monograph No. 7). February, 1979.

Foster, G., Algozzine, B., & Ysseldyke, J. Susceptibility to stereotypic bias (Research Report No. 3). March, 1979.

Algozzine, B. An analysis of the disturbingness and acceptability of behaviors as a function of diagnostic label (Research Report No. 4). March, 1979.

Algozzine, B., & McGraw, K. Diagnostic testing in mathematics: An extension of the PIAT? (Research Report No. 5). March, 1979.

Deno, S. L. A direct observation approach to measuring classroom behavior: Procedures and application (Research Report No. 6). April, 1979.

Ysseldyke, J. E., & Mirkin, P. K. Proceedings of the Minnesota roundtable conference on assessment of learning disabled children (Monograph No. 8). April, 1979.

Somwaru, J. P. A new approach to the assessment of learning disabilities (Monograph No. 9). April, 1979.

Algozzine, B., Forgnone, C., Mercer, C. D., & Trifiletti, J. J. Toward defining discrepancies for specific learning disabilities: An analysis and alternatives (Research Report No. 7). June, 1979.

Algozzine, B. The disturbing child: A validation report (Research Report No. 8). June, 1979.

Note: Monographs No. 1 - 6 and Research Report No. 2 are not available for distribution. These documents were part of the Institute's 1979-1980 continuation proposal, and/or are out of print.

Ysseldyke, J. E., Algozzine, B., Regan, R., & Potter, M. Technical adequacy of tests used by professionals in simulated decision making (Research Report No. 9). July, 1979.

Jenkins, J. R., Deno, S. L., & Mirkin, P. K. Measuring pupil progress toward the least restrictive environment (Monograph No. 10). August, 1979.

Mirkin, P. K., & Deno, S. L. Formative evaluation in the classroom: An approach to improving instruction (Research Report No. 10). August, 1979.

Thurlow, M. L., & Ysseldyke, J. E. Current assessment and decision-making practices in model programs for the learning disabled (Research Report No. 11). August, 1979.

Deno, S. L., Chiang, B., Tindal, G., & Blackburn, M. Experimental analysis of program components: An approach to research in CSDC's (Research Report No. 12). August, 1979.

Ysseldyke, J. E., Algozzine, B., Shinn, M., & McGue, M. Similarities and differences between underachievers and students labeled learning disabled: Identical twins with different mothers (Research Report No. 13). September, 1979.

Ysseldyke, J., & Algozzine, R. Perspectives on assessment of learning disabled students (Monograph No. 11). October, 1979.

Poland, S. F., Ysseldyke, J. E., Thurlow, M. L., & Mirkin, P. K. Current assessment and decision-making practices in school settings as reported by directors of special education (Research Report No. 14). November, 1979.

McGue, M., Shinn, M., & Ysseldyke, J. Validity of the Woodcock-Johnson psycho-educational battery with learning disabled students (Research Report No. 15). November, 1979.

Deno, S., Mirkin, P., & Shinn, M. Behavioral perspectives on the assessment of learning disabled children (Monograph No. 12). November, 1979.

Sutherland, J. H., Algozzine, B., Ysseldyke, J. E., & Young, S. What can I say after I say LD? (Research Report No. 16). December, 1979.

Deno, S. L., & Mirkin, P. K. Data-based IEP development: An approach to substantive compliance (Monograph No. 13). December, 1979.

Ysseldyke, J., Algozzine, B., Regan, R., & McGue, M. The influence of test scores and naturally-occurring pupil characteristics on psycho-educational decision making with children (Research Report No. 17). December, 1979.

Algozzine, B., & Ysseldyke, J. E. Decision makers' prediction of students' academic difficulties as a function of referral information (Research Report No. 18). December, 1979.

Ysseldyke, J. E., & Algozzine, B. Diagnostic classification decisions as a function of referral information (Research Report No. 19). January, 1980.

Deno, S. L., Mirkin, P. K., Chiang, B., & Lowry, L. Relationships among simple measures of reading and performance on standardized achievement tests (Research Report No. 20). January, 1980.

Deno, S. L., Mirkin, P. K., Lowry, L., & Kuehnle, K. Relationships among simple measures of spelling and performance on standardized achievement tests (Research Report No. 21). January, 1980.

Deno, S. L., Mirkin, P. K., & Marston, D. Relationships among simple measures of written expression and performance on standardized achievement tests (Research Report No. 22). January, 1980.

Mirkin, P. K., Deno, S. L., Tindal, G., & Kuehnle, K. Formative evaluation: Continued development of data utilization systems (Research Report No. 23). January, 1980.

Deno, S. L., Mirkin, P. K., Robinson, S., & Evans, P. Relationships among classroom observations of social adjustment and sociometric rating scales (Research Report No. 24). January, 1980.

Thurlow, M. L., & Ysseldyke, J. E. Factors influential on the psychoeducational decisions reached by teams of educators (Research Report No. 25). February, 1980.

Ysseldyke, J. E., & Algozzine, B. Diagnostic decision making in individuals susceptible to biasing information presented in the referral case folder (Research Report No. 26). March, 1980.

Thurlow, M. L., & Greener, J. W. Preliminary evidence on information considered useful in instructional planning (Research Report No. 27). March, 1980.

Ysseldyke, J. E., Regan, R. R., & Schwartz, S. Z. The use of technically adequate tests in psychoeducational decision making (Research Report No. 28). April, 1980.

Richey, L., Potter, M., & Ysseldyke, J. Teachers' expectations for the siblings of learning disabled and non-learning disabled students: A pilot study (Research Report No. 29). May, 1980.

Thurlow, M. L., & Ysseldyke, J. E. Instructional planning: Information collected by school psychologists vs. information considered useful by teachers (Research Report No. 30). June, 1980.

Algozzine, B., Webber, J., Campbell, M., Moore, S., & Gilliam, J. Classroom decision making as a function of diagnostic labels and perceived competence (Research Report No. 31). June, 1980.

Ysseldyke, J. E., Algozzine, B., Regan, R. R., Potter, M., Richey, L., & Thurlow, M. L. Psychoeducational assessment and decision making: A computer-simulated investigation (Research Report No. 32). July, 1980.

Ysseldyke, J. E., Algozzine, B., Regan, R. R., Potter, M., & Richey, L. Psychoeducational assessment and decision making: Individual case studies (Research Report No. 33). July, 1980.

Ysseldyke, J. E., Algozzine, B., Regan, R., Potter, M., & Richey, L. Technical supplement for computer-simulated investigations of the psychoeducational assessment and decision-making process (Research Report No. 34). July, 1980.

Algozzine, B., Stevens, L., Costello, C., Beattie, J., & Schmid, R. Classroom perspectives of LD and other special education teachers (Research Report No. 35). July, 1980.

Algozzine, B., Siders, J., Siders, J., & Beattie, J. Using assessment information to plan reading instructional programs: Error analysis and word attack skills (Monograph No. 14). July, 1980.

Ysseldyke, J., Shinn, M., & Epps, S. A comparison of the WISC-R and the Woodcock-Johnson Tests of Cognitive Ability (Research Report No. 36). July, 1980.

Algozzine, B., & Ysseldyke, J. E. An analysis of difference score reliabilities on three measures with a sample of low achieving youngsters (Research Report No. 37). August, 1980.

Shinn, M., Algozzine, B., Marston, D., & Ysseldyke, J. A theoretical analysis of the performance of learning disabled students on the Woodcock-Johnson Psycho-Educational Battery (Research Report No. 38). August, 1980.

Richey, L. S., Ysseldyke, J., Potter, M., Regan, R. R., & Greener, J. Teachers' attitudes and expectations for siblings of learning disabled children (Research Report No. 39). August, 1980.

Ysseldyke, J. E., Algozzine, B., & Thurlow, M. L. (Eds.). A naturalistic investigation of special education team meetings (Research Report No. 40). August, 1980.

Meyers, B., Meyers, J., & Deno, S. Formative evaluation and teacher decision making: A follow-up investigation (Research Report No. 41). September, 1980.

Fuchs, D., Garwick, D. R., Featherstone, N., & Fuchs, L. S. On the determinants and prediction of handicapped children's differential test performance with familiar and unfamiliar examiners (Research Report No. 42). September, 1980.

Algozzine, B., & Stoller, L. Effects of labels and competence on teachers' attributions for a student (Research Report No. 43). September, 1980.

Ysseldyke, J. E., & Thurlow, M. L. (Eds.). The special education assessment and decision-making process: Seven case studies (Research Report No. 44). September, 1980.

Ysseldyke, J. E., Algozzine, B., Potter, M., & Regan, R. A descriptive study of students enrolled in a program for the severely learning disabled (Research Report No. 45). September, 1980.

Marston, D. Analysis of subtest scatter on the tests of cognitive ability from the Woodcock-Johnson Psycho-Educational Battery (Research Report No. 46). October, 1980.

Algozzine, B., Ysseldyke, J. E., & Shinn, M. Identifying children with learning disabilities: When is a discrepancy severe? (Research Report No. 47). November, 1980.

Fuchs, L., Tindal, J., & Deno, S. Effects of varying item domain and sample duration on technical characteristics of daily measures in reading (Research Report No. 48). January, 1981.

Marston, D., Lowry, L., Deno, S., & Mirkin, P. An analysis of learning trends in simple measures of reading, spelling, and written expression: A longitudinal study (Research Report No. 49). January, 1981.

Marston, D., & Deno, S. The reliability of simple, direct measures of written expression (Research Report No. 50). January, 1981.

Epps, S., McGue, M., & Ysseldyke, J. E. Inter-judge agreement in classifying students as learning disabled (Research Report No. 51). February, 1981.

Epps, S., Ysseldyke, J. E., & McGue, M. Differentiating LD and non-LD students: "I know one when I see one" (Research Report No. 52). March, 1981.

Evans, P. R., & Peham, M. A. S. Testing and measurement in occupational therapy. A review of current practice with special emphasis on the Southern California Sensory Integration Tests (Monograph No. 15). April, 1981.

Fuchs, L., Wesson, C., Tindal, G., & Mirkin, P. Teacher efficiency in continuous evaluation of IEP goals (Research Report No. 53). June, 1981.

Fuchs, D., Featherstone, N., Garwick, D. R., & Fuchs, L. S. The importance of situational factors and task demands to handicapped children's test performance (Research Report No. 54). June, 1981.

Tindal, G., & Deno, S. L. Daily measurement of reading: Effects of varying the size of the item pool (Research Report No. 55). July, 1981.

Fuchs, L. S., & Deno, S. L. A comparison of teacher judgment, standardized tests, and curriculum-based approaches to reading placement (Research Report No. 56). August, 1981.

Fuchs, L., & Deno, S. The relationship between curriculum-based mastery measures and standardized achievement tests in reading (Research Report No. 57). August, 1981.

Christenson, S., Graden, J., Potter, M., & Ysseldyke, J. Current research on psychoeducational assessment and decision making: Implications for training and practice (Monograph No. 16). September, 1981.

Christenson, S., Ysseldyke, J., & Algozzine, B. Institutional constraints and external pressures influencing referral decisions (Research Report No. 58). October, 1981.

Fuchs, L., Fuchs, D., & Deno, S. Reliability and validity of curriculum-based informal reading inventories (Research Report No. 59). October, 1981.

Algozzine, B., Christenson, S., & Ysseldyke, J. Probabilities associated with the referral-to-placement process (Research Report No. 60). November, 1981.

Tindal, G., Fuchs, L. Christenson, S., Mirkin, P., & Deno, S. The relationship between student achievement and teacher assessment of short- or long-term goals (Research Report No. 61). November, 1981.

Mirkin, P., Fuchs, L., Tindal, G., Christenson, S., & Deno, S. The effect of IEP monitoring strategies on teacher behavior (Research Report No. 62). December, 1981.

Wesson, C., Mirkin, P., & Deno, S. Teachers' use of self instructional materials for learning procedures for developing and monitoring progress on IEP goals (Research Report No. 63). January, 1982.

Fuchs, L., Wesson, C., Tindal, G., Mirkin, P., & Deno, S. Instructional changes, student performance, and teacher preferences: The effects of specific measurement and evaluation procedures (Research Report No. 64). January, 1982.

Potter, M., & Mirkin, P. Instructional planning and implementation practices of elementary and secondary resource room teachers: Is there a difference? (Research Report No. 65). January, 1982.

Thurlow, M. L., & Ysseldyke, J. E. Teachers' beliefs about LD students (Research Report No. 66). January, 1982.

Graden, J., Thurlow, M. L., & Ysseldyke, J. E. Academic engaged time and its relationship to learning: A review of the literature (Monograph No. 17). January, 1982.

King, R., Wesson, C., & Deno, S. Direct and frequent measurement of student performance: Does it take too much time? (Research Report No. 67). February, 1982.

Greener, J. W., & Thurlow, M. L. Teacher opinions about professional education training programs (Research Report No. 68). March, 1982.

Algozzine, B., & Ysseldyke, J. Learning disabilities as a subset of school failure: The oversophistication of a concept (Research Report No. 69). March, 1982.

Fuchs, D., Zern, D. S., & Fuchs, L. S. A microanalysis of participant behavior in familiar and unfamiliar test conditions (Research Report No. 70). March, 1982.

Shinn, M. R., Ysseldyke, J., Deno, S., & Tindal, G. A comparison of psychometric and functional differences between students labeled learning disabled and low achieving (Research Report No. 71). March, 1982.

Thurlow, M. L. Graden, J., Greener, J. W., & Ysseldyke, J. E. Academic responding time for LD and non-LD students (Research Report No. 72). April, 1982.

Graden, J., Thurlow, M., & Ysseldyke, J. Instructional ecology and academic responding time for students at three levels of teacher-perceived behavioral competence (Research Report No. 73). April, 1982.

Algozzine, B., Ysseldyke, J., & Christenson, S. The influence of teachers' tolerances for specific kinds of behaviors on their ratings of a third grade student (Research Report No. 74). April, 1982.

Wesson, C., Deno, S., & Mirkin, P. Research on developing and monitoring progress on IEP goals: Current findings and implications for practice (Monograph No. 18). April, 1982.

Mirkin, P., Marston, D., & Deno, S. L. Direct and repeated measurement of academic skills: An alternative to traditional screening, referral, and identification of learning disabled students (Research Report No. 75). May, 1982.

Algozzine, B., Ysseldyke, J., Christenson, S., & Thurlow, M. Teachers' intervention choices for children exhibiting different behaviors in school (Research Report No. 76). June, 1982.

Tucker, J., Stevens, L. J., & Ysseldyke, J. E. Learning disabilities: The experts speak out (Research Report No. 77). June, 1982.

Thurlow, M. L., Ysseldyke, J. E., Graden, J., Greener, J. W., & Mecklenberg, C. Academic responding time for LD students receiving different levels of special education services (Research Report No. 78). June, 1982.

Graden, J. L., Thurlow, M. L., Ysseldyke, J. E., & Algozzine, B. Instructional ecology and academic responding time for students in different reading groups (Research Report No. 79). July, 1982.

Mirkin, P. K., & Potter, M. L. A survey of program planning and implementation practices of LD teachers (Research Report No. 80). July, 1982.

Fuchs, L. S., Fuchs, D., & Warren, L. M. Special education practice in evaluating student progress toward goals (Research Report No. 81). July, 1982.

Kuehnle, K., Deno, S. L., & Mirkin, P. K. Behavioral measurement of social adjustment: What behaviors? What setting? (Research Report No. 82). July, 1982.

Fuchs, D., Dailey, Ann Madsen, & Fuchs, L. S. Examiner familiarity and the relation between qualitative and quantitative indices of expressive language (Research Report No. 83). July, 1982.

Videen, J., Deno, S., & Marston, D. Correct word sequences: A valid indicator of proficiency in written expression (Research Report No. 84). July, 1982.

Potter, M. L. Application of a decision theory model to eligibility and classification decisions in special education (Research Report No. 85). July, 1982.

Greener, J. E., Thurlow, M. L., Graden, J. L., & Ysseldyke, J. E. The educational environment and students' responding times as a function of students' teacher-perceived academic competence (Research Report No. 86). August, 1982.

Deno, S., Marston, D., Mirkin, P., Lowry, L., Sindelar, P., & Jenkins, J. The use of standard tasks to measure achievement in reading, spelling, and written expression: A normative and developmental study (Research Report No. 87). August, 1982.

Skiba, R., Wesson, C., & Deno, S. L. The effects of training teachers in the use of formative evaluation in reading: An experimental-control comparison (Research Report No. 88). September, 1982.

Marston, D., Tindal, G., & Deno, S. L. Eligibility for learning disability services: A direct and repeated measurement approach (Research Report No. 89). September, 1982.

Thurlow, M. L., Ysseldyke, J. E., & Graden, J. L. LD students' active academic responding in regular and resource classrooms (Research Report No. 90). September, 1982.

Ysseldyke, J. E., Christenson, S., Pianta, R., Thurlow, M. L., & Algozzine, B. An analysis of current practice in referring students for psychoeducational evaluation: Implications for change (Research Report No. 91). October, 1982.

Ysseldyke, J. E., Algozzine, B., & Epps, S. A logical and empirical analysis of current practices in classifying students as handicapped (Research Report No. 92). October, 1982.

Tindal, G., Marston, D., Deno, S. L., & Germann, G. Curriculum differences in direct repeated measures of reading (Research Report No. 93). October, 1982.

Fuchs, L.S., Deno, S. L., & Marston, D. Use of aggregation to improve the reliability of simple direct measures of academic performance (Research Report No. 94). October, 1982.

Ysseldyke, J. E., Thurlow, M. L., Mecklenburg, C., & Graden, J. Observed changes in instruction and student responding as a function of referral and special education placement (Research Report No. 95). October, 1982.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. Effects of frequent curriculum-based measurement and evaluation on student achievement and knowledge of performance: An experimental study (Research Report No. 96). November, 1982.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. Direct and frequent measurement and evaluation: Effects on instruction and estimates of student progress (Research Report No. 97). November, 1982.

Tindal, G., Wesson, C., Germann, G., Deno, S. L., & Mirkin, P. K. The Pine County model for special education delivery: A data-based system (Monograph No. 19). November, 1982.

Epps, S., Ysseldyke, J. E., & Algozzine, B. An analysis of the conceptual framework underlying definitions of learning disabilities (Research Report No. 98). November, 1982.

Epps, S., Ysseldyke, J. E., & Algozzine, B. Public-policy implications of different definitions of learning disabilities (Research Report No. 99). November, 1982.

Ysseldyke, J. E., Thurlow, M. L., Graden, J. L., Wesson, C., Deno, S. L., & Algozzine, B. Generalizations from five years of research on assessment and decision making (Research Report No. 100). November, 1982.

Marston, D., & Deno, S. L. Measuring academic progress of students with learning difficulties: A comparison of the semi-logarithmic chart and equal interval graph paper (Research Report No. 101). November, 1982.

Beattie, S., Grise, P., & Algozzine, B. Effects of test modifications on minimum competency test performance of third grade learning disabled students (Research Report No. 102). December, 1982

Algozzine, B., Ysseldyke, J. E., & Christenson, S. An analysis of the incidence of special class placement: The masses are burgeoning (Research Report No. 103). December, 1982.

Marston, D., Tindal, G., & Deno, S. L. Predictive efficiency of direct, repeated measurement: An analysis of cost and accuracy in classification (Research Report No. 104). December, 1982.

Wesson, C., Deno, S., Mirkin, P., Sevcik, B., Skiba, R., King, R., Tindal, G., & Maruyama, G. Teaching structure and student achievement effects of curriculum-based measurement: A causal (structural) analysis (Research Report No. 105). December, 1982.

Mirkin, P. K., Fuchs, L. S., & Deno, S. L. (Eds.). Considerations for designing a continuous evaluation system: An integrative review (Monograph No. 20). December, 1982.